The Dissertation Committee for Azat Akhmetov
certifies that this is the approved version of the following dissertation:

Yeast as a Platform For Synthetic Biology and Investigation of Evolutionary Hypotheses

Committee:

Edward M. Marcotte, Supervisor

Hal S. Alper

Andrew D. Ellington

Makkuni Jayaram

Claus O. Wilke

Yeast as a Platform For Synthetic Biology and Investigation of Evolutionary Hypotheses

by

Azat Akhmetov

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2019

I dedicate this thesis to my lovely, loving wife Larissa. Without her, this humble text would not have been possible, nor would a great many wonderful things.

Acknowledgments

I would like to express my infinite gratitude to my advisor, Edward Marcotte. Without his support, encouragement and guidance my research would have surely been doomed to failure. He has been an ocean of kindness, patience, compassion and understanding during the many ups and downs of any research project's course. When I first came to UT, I was convinced that surely I knew everything. As Edward's student, I have learned so much about biology, statistics, data and science that I cannot imagine how I managed to think I knew anything.

I would also like to thank my wonderful colleagues, past and present, at the Marcotte Lab and the UT community at large, for invaluable discussion and feedback on my research. In particular, I acknowledge:

- Aashiq and Jon Laurent with whom we've had many deep discussions about yeast genomics, brainstormed fantastic project ideas (some even became actual projects!) and collaborated on several papers.
- Chris Yellman, who has illuminated my mind to such intricacies of the life histories of yeasts as had thwarted my imagination and comprehension both. Also, the yeasts themselves mostly came from him.
- Madelyn, who was possibly the most brilliant undergraduate I worked with.

- Jag, Alex and Riddhiman who have directly inspired much of my research and gave me so many great ideas.

I am also deeply grateful to my committee members Andy Ellington, Claus Wilke, Jayaram Makkuni and Hal Alper for their patience and valuable feedback on my research. I'd like to thank Andy especially for his brilliant contributions to our DNA archival paper.

Last but not least, I would like to thank my parents, Nail and Gulshat, who have who have always supported me and driven me to succeed while letting me find my own path through life. I'd also like to express my love and gratitude to Larissa, my better half, for enduring for so long my bizarre grad student lifestyle.

Yeast as a Platform For Synthetic Biology and Investigation of Evolutionary Hypotheses

Azat Akhmetov, Ph.D.
The University of Texas at Austin, 2019

Supervisor: Edward M. Marcotte

Yeast has long been the *sine qua non* of model organisms due to its experimental tractability. Recent advances in biology, such as CRISPR/Cas9 editing, promise to bring this tractability to other model organism. But the same advances have also added new dimensions to the utility of yeast, and reinforced its importance as a unique platform for studying basic biology, translational research and building the future of biotechnology. In this thesis I describe (i) the use of humanized yeast to study fundamental questions about the evolution of genetic systems, (ii) rapid and versatile techniques for leveraging classical yeast techniques along with cutting edge developments to solve 21-st century problems and (iii) a blueprint for transforming yeast into a vessel for storage of digital information.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Since the earliest days of molecular biology, the baker's yeast *Saccharomyces cerevisiae* has enjoyed great prominence as a platform for synthetic biology and bioengineering (indeed, humanity has used yeast for many millennia prior to the advent of biology[1]). As a unicellular organism, it can grow prodigiously in laboratory cultures. Like many fungi, it has an excellent capacity for artificial biosynthesis. As a eukaryote, it is often a much more convenient vehicle for expressing systems derived from important multicellular organisms, such as humans or crops. But last if not least, yeast has a robust DNA repair system – as a result, countless new techniques of modifying DNA have been pioneered in this organism. Many are still vastly more practical in it to this day. The research presented in this thesis represents my own efforts at expanding the capabilities of this powerful biotechnology platform even further, to aid investigation of theory and practical technology alike.

1.1 Genetic Modification in Yeast

Even before modern biology, humanity modified the yeast genome through selective breeding[2]. Likely, this modification continued unwittingly

for many centuries before humanity even became aware that microorganisms exist[3]. Likewise, since the dawn of the modern era, it was very actively studied[4]. As such, techniques of yeast modification are legion, and I trust the reader to kindly excuse me for eschewing to attempt a comprehensive inventory of them in this dissertation.

Yeast genetic engineering was already quite advanced before the advent of CRISPR. The state of the art was the *delitto perfetto*[1] method[5]. This method joined two families of genome editing techniques to provide the ultimate editing method: Before, there were methods that could make any edit, but only at certain loci. For instance, the I-SceI endonuclease could cleave DNA effectively, which triggered homologous recombination (HR) and could be used to incorporate a new sequence replacing the old one. But I-SceI targets only a specific sequence. There were also methods that could edit any locus, but the efficiency was low and required selection, which in turn required introducing a selectable marker in the edit. In *delitto perfetto*, the latter approach is used to first introduce a cassette into the target site. Clones are isolated by selection. Besides a selectable marker, the cassette carries a recognition site for I-SceI, which is then expressed in the clones. This leads to

[1] This method is remarkable in that it is not just brilliantly conceived, but also brilliantly named. The term is Italian for "perfect murder". One of the chief uses of genome editing techniques is to delete sequence (such as genes of interest), but prior techniques would always introduce some extraneous sequence at the site of editing (the so-called "scar"). Correctly applied, the *delitto perfetto* leaves no scar. It was, therefore, a qualitative improvement because it could edit so as to generate *any* desired sequence. Other methods could not; they only produced sequences which include the scar. Hence *delitto perfetto* could delete a gene, leaving behind no evidence of the deed.

a second round of HR which replaces the cassette with the desired sequence.

While this method is powerful, it has two weaknesses: Firstly, the two rounds of growth and selection are somewhat laborious in routine use. Second, after the first part, the cassette will potentially interrupt the function of genomic sequence it is targeting. If that sequence is an essential gene, the clone will be inviable. Therefore, an additional layer of complexity must be introduced in the form of a temporary complementor gene, such as a plasmid carrying a functional version of the gene. However, when HR is triggered the cassette site will be edited to incorporate a homologous sequence. If the complementing plasmid carries the yeast gene, it will be highly homologous and may even outcompete the repair template of interest. Therefore the complementing plasmid must be carefully engineered to be similar enough to rescue disruption by the cassette, but different enough to not be picked up by HR. In many cases, these issues are quite tolerable. However, in more complex and sophisticated studies they can seriously complicate experimental design. They were a direct motivation for me to develop the CRISPR-based method described in Chapter 3, which resolved them and enabled the work of my other chapters.

Nevertheless, the techniques available in yeast were often far ahead of other organisms, and yeast itself if very easy to culture and work with. Large-scale application of the *delitto perfetto* method and others produced large datasets and collections of mutant strains. Genome-wide knockout libraries revealed foundational insights of functional genomics[6]. This deletion

library was then used to probe the relationship between genotype, phenotype, and environment with unprecedented depth and breadth[7]. This deletion library was also the basis for our lab's initial systematic humanization study[8], which served as the foundation of my work. Mating such libraries allowed researchers to explore combinatorial effects on a vast scale. Earlier, the yeast-2-hybrid technique had elucidated protein-protein interaction networks[9, 10]. Now the development of synthetic genetic arrays[11] elucidated gene-gene interactions[12].

1.1.1 CRISPR/Cas9 Editing

CRISPR is a system of bacterial adaptive immunity[13, 14]. Many bacteria contain short segments of DNA in their genomes, which match sequences of common foreign DNA such as phage genomes. These segments are transcribed as RNA, which forms a complex with other CRISPR proteins such as Cas9. The RNA is able to insert itself into double-stranded DNA (dsDNA) and form base pairs. Upon successful base pairing, which effectively serves as target confirmation, the protein complex begins cutting the DNA with its endonuclease activity. In this way, bacteria in nature can selectively destroy a broad spectrum of harmful foreign DNA. The complementary RNA sequence is quite short (classically 20 bp) and determines the target by matching its sequence. Accordingly, it is called the guide RNA. There is an additional RNA component which does not vary between targets but is necessary for CRISPR function, and it is termed the scaffold RNA.

It did not take long for the scientific community to perceive the utility of the CRISPR system in biotechnology applications. In many organisms, genome editing is accomplished by relying on native damage repair enzymes to incorporate a desired sequence into the genome. For this to be a targeted process (which is required for true editing, rather than mere insertion), there must be a way of preferentially inducing DNA damage at the targeted locus. Enzymes capable of targeted DNA disruption have been known since the earliest days of molecular biology[15]. However, in many cases, each enzyme could target only a small range of sequences. Accordingly, biologists have resorted to amassing extensive libraries of enzymes for different sequences. If no enzyme was available for a given site, it was very difficult to modify it. Moreover, restriction enzymes were often short enough to be prolific throughout the genome, precluding specific nuclease activity at the target site only. Programmable nucleases have attracted much research interest as a result[16, 17]. Soon after the native function of the CRISPR system was understood, a now-famous study showed that CRISPR can be used as just such a programmable nuclease[18]. Because CRISPR targeting is controlled by base pairing of the guide RNA, any sequence can be targeted by simply inserting a complementary sequence into the guide. Because the guide sequence is about 20 bp long, it has a very low probability of occurring in more than one place (with the important exception of duplicate genes). They also fused many components of the system, reducing it to a single protein (SpyCas9, derived from *Staphylococcus pyogenes*) and a single RNA (called sgRNA, being a fusion of

scaffold and guide) which can easily be expressed in most cells. Soon after, systems were developed to further enhance the precision of CRISPR targeting by increasing the number of guide basepairs that participate[19].

The earliest application of CRISPR to yeast was by DiCarlo and colleagues and influenced much of subsequent work[20]. They cloned the system from [18] into yeast vectors and demonstrated proof of concept experiments for use of CRISPR to edit yeast DNA (in conjunction with a repair template) or mutagenize a specific site (by omitting the repair template). As it turned out, yeast was exceptionally good at repairing double-stranded breaks by homologous recombination. This allowed very straightforward protocols for making arbitrary edits to the yeast genome. CRISPR editing in yeast was so efficient, that it obviated the need for selection, permitting single step genome modification[21]. A minor problem was expressing sufficient amounts of sgRNA: Initially, its expression was driven by small nuclear RNA promoters which were not very strong. This problem was soon remedied by introducing stronger promoters[22]. Another challenge was rapid cloning of many CRISPR guides, also resolved shortly[23].

In the context of yeast, the principle of CRISPR editing is fairly simple: One must express the modified Cas9 protein and the appropriate sgRNA, and cleavage will usually occur at a high rate. Both can be driven from a single plasmid. Cas9 is modified with a nuclear localization signal to deliver it to the nucleus. Although it is sizable (the coding sequence is about 5 kb long), expressing it rarely poses problems. Guide RNA design is in principle simple,

but choice of target is not. Although CRISPR activity is usually efficient and specific, it is known to have a complex dependence on the exact sequence context of the target[24]. Fortunately, the CRISPR site need not be exactly at the target – induction of double-strand break in relative proximity of the desired replacement is often sufficient. Because of this, it is worthwhile to hedge by generating multiple guides for every genomic target and empirically testing their activity. This is the route I have taken in my own research.

1.2 Cross-species Ortholog Complementation in Yeast

Humanization in a broad sense of simply introducing human genes into yeast is an established research strategy. It was instrumental in revealing the biology of key pathways[25] and proteins such as RAS[26], CDC2[27] and PIF1[28]. Systematic, genome-scale humanization studies have opened new horizons for this field and attracted a great deal of interest[8]. In this publication, yeasts from a deletion library were complemented with human genes (Figure 1.2). The study focused on human orthologs of essential yeast genes — with the surprising finding that in roughly half of the cases the yeast gene can be swapped with its human ortholog.

When a cross-species ortholog is complementing an essential function, disrupting that ortholog should logically bring about a phenotype corresponding to disruption of that function. The phenotype of a mutant complementing ortholog might be the same as that for mutations of the native yeast gene, or it may be a novel phenotype. Regardless, the mere presence of the phenotype

is sufficient to detect disruption of function. For example, dysfunctional mutants of human genes can be compared to their wild types on the basis of their phenotype in humanized yeast, even if the phenotype does not match or even resemble that of the dysfunctional mutant in humans. This recalls the concept of phenologs[30, 29] wherein conserved and related genetic modules, many of their constituents orthologous will also have conserved molecular responses to genetic disruption. This response may be manifested in very different ways according to the physiology of each organism. However, disruption of a conserved genetic element in a pair of species would similarly induce a phenotype for both species, even if it is not the same phenotype. Phenologs have been used to identify human drugs in the context of very dissimilar organisms[31]. Humanoid yeast can be regarded as a case of augmented phenologs, where the introduction of human genes enables a higher degree of correspondence to human physiology. Humanized genes in yeast would not generate the same phenotype as in humans, but mutations and active drugs identified through screens for phenotypes in yeast would be high-value targets to investigate for the same activity in humans[32]. There have been recent pioneering studies demonstrating that humanized yeast is a viable platform for screening human alleles[33]. Such humanization assays can identify deleterious mutant alleles with considerably better performance than computational methods, while still recapitulating many computational findings [34]. There are thousands of yeast-human orthologs[35]; but in cases where suitable ortholog pairs are not available, it is still feasible to apply humanization assays via paralogs[36].

1.3 This Thesis

In Chapter 2, I will present research involving systematic swapping of several yeast genes with their orthologs from humans, bacteria, plants, and archea[37]. This research yielded interesting insights into evolution (such as localization of metabolic activity) and also expanded on the general results of a similar earlier study[8]. The work described here gives particular attention to the heme biosynthesis pathway in yeast, which is has a very universal role across the tree of life. In yeast, it participates in respiration and maintenance of the cellular envelope. In humans, orthologs of these genes are involved in the production of hemoglobin and are implicated in porphyria disease. In plants, they are involved in chlorophyll production and are frequently targeted by pesticides. An interesting finding was that the bacterial orthologs mislocalize in yeast and cause significant impairment, with a phenotype highly reminiscent of human porphyria. This serendipitous discovery is an excellent demonstration of how ortholog swapping in yeast can help us understand the phenotype-genotype relationship even in organisms very distantly related to yeast.

In Chapter 3 I present a detailed protocol for CRISPR-mediated yeast humanization and genome editing, which was developed alongside the work in Chapter 2 and contributed to that work as well. The exciting findings we observed for heme are only the beginning: There is a wealth of biology to be learned from expanding this approach to more pathways and complexes. My hope is that the protocol we developed will serve as a foundation for ortholog

swapping experiments on a massive scale.

In Chapter 4 I describe my latest work in building upon and expanding on the two previous chapters. I have endeavored to find the limit of humanization of the heme pathway and explore the idea of whole-pathway humanization. I describe using these humanized strains to survey epistasis and evolution of genetic modules.

Chapter 5 takes a different turn into the interface of digital storage and biotechnology. Here I propose a system for encoding digital data as DNA which can be stored in living cells such as yeast. With the invention of computers, electronic data storage rapidly displaced paper due to its many advantages. However, it has a significant disadvantage: Longevity. Conversely, billions of years of evolution have molded DNA into the ultimate information store. It also enables storage densities higher than what is possible with digital circuits. Lastly, every new storage technology necessarily requires the establishment of new infrastructure[38]. DNA is the exception, the tools for reading and writing it already exist, and will conceivably continue to exist so long as civilization endures. Even after our civilization ends, any future civilization will almost certainly rediscover them. In this sense, DNA seems a poetic solution to the data storage problem. This work also continues my theme of exploring the frontiers of synthetic biology in the 21st century.

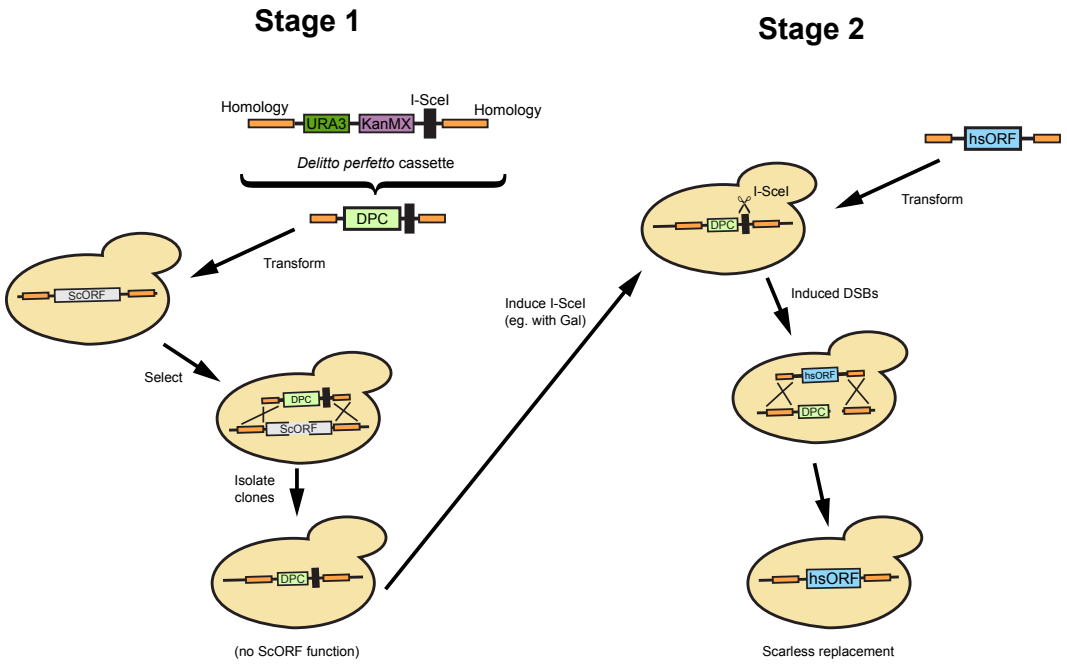

Figure 1.1: The *delitto perfetto* method of yeast genome editing[5]. (Continued on next page.)

Figure 1.1: The *delitto perfetto* method is a two stage process. First, homologies to the target region of the genome are added to a selectable cassette. The cassette is transformed into yeast cells, and selected for. Yeast has a very low background rate of DNA breakage, on the order of 10^{-6}. When a large number of cells is transformed, DNA breakage will appear close to the target site in some cells, and its repair will result in incorporation of the cassette. Though these cells are a tiny minority, they can be selected via the marker in the cassette (e.g. URA3, selected for with -Ura medium). The marker also carries a recognition site for the I-SceI endonuclease. Once clones carrying the cassette have been isolated, they can be taken through the second step: I-SceI is expressed (e.g. from a plasmid) while the actual repair template carrying the desired sequence is introduced. I-SceI begins to cleave the genome at the cassette, which triggers the DNA repair once again, and this time results in replacement of the cassette with the desired sequence. Because URA3 is counter-selectable by culture on 5-FOA, its elimination can also be selected for.

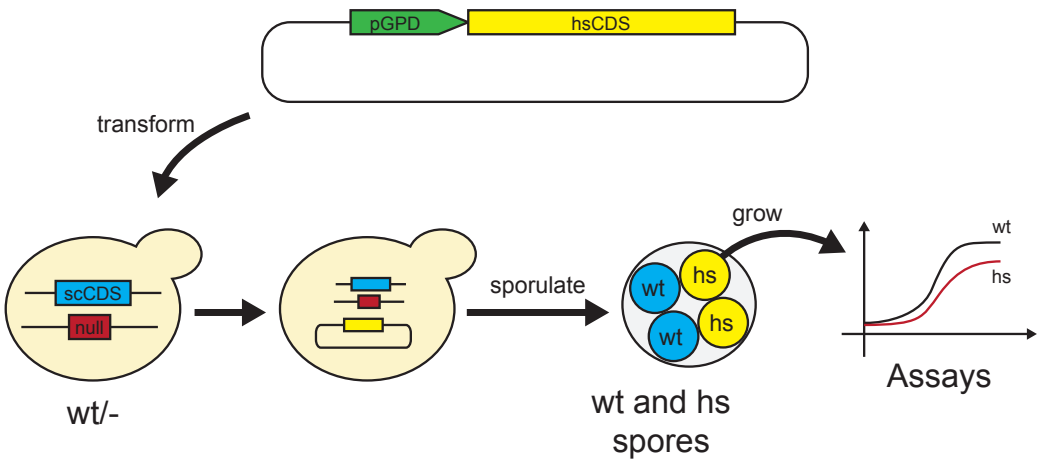

Figure 1.2: Overview of the humanization method in [8]. In this study, heterozygous deletion strains of essential yeast genes with 1:1 human orthologs were obtained from a deletion library. The human ortholog, under control of a constitutive promoter, was introduced into these yeasts. Afterwards, they were sporulated and tetrads were dissected to obtain haploid cells which have the null allele. Because the gene is essential, all 4 spores can be obtained only if the human gene was able to complement the deletion. If the gene was not replaceable, only 2 spores (with the wild type allele) would be obtained.

Chapter 2

Systematic Bacterialization of Yeast Genes Identifies a Near-universally Swappable Pathway

After the surprising discovery that a large number of yeast genes can be successfully humanized, I became interested in the extent to which complementation is still viable across the tree of life. To this end, I collaborated with Aashiq Kachroo and Jon Laurent on a project where we used genomic editing to attempt replacement of yeast genes with not just human, but also plant and bacterial orthologs.

The biosynthesis of heme stood out as a striking example of universal replaceability. This pathway is known for its role in human biology because it produces a porphyrin molecule, which is complexed to iron in the final step of the pathway and becomes a component of hemoglobin. However, heme itself also has other roles, such as participating in the cellular respiration as an electron acceptor, which is why it is found in many other organisms that do not have blood, like yeast or bacteria. In yeast, the pathway is also neces-

This chapter was previously published in Kachroo AH, Laurent JM, Akhmetov A, *et al.* (2017) *Elife*, 6:e25093. I contributed to conducting the initial replacement experiments, validation of the results, follow up work, analysis, and writing.

sary for the synthesis of ergosterol, which is unique to fungi and an essential component of their cellular membrane. In plants, one of the intermediates in this pathway is the precursor of chlorophyll synthesis. In sum, heme and therefore its biosynthesis, have a central and fundamental role in almost all living organisms. It also has similarly critical secondary roles across kingdoms and domains of life. Because of this, that it can be swapped at all between organisms as divergent as *S. cerevisiae*, *E. coli*, *A. thaliana* and *H. sapiens* is in itself startling. But furthermore, in-depth analysis of these swaps can shed new light on key evolutionary questions. This is elaborated in detail in a report describing our work, which was published in the journal eLife[37].

2.1 Abstract

Eukaryotes and prokaryotes last shared a common ancestor 2 billion years ago, and while many present-day genes in these lineages predate this divergence, the extent to which these genes still perform their ancestral functions is largely unknown. To test principles governing retention of ancient function, we asked if prokaryotic genes could replace their essential eukaryotic orthologs. We systematically replaced essential genes in yeast by their 1:1 orthologs from Escherichia coli. After accounting for mitochondrial localization and alternative start codons, 31 out of 51 bacterial genes tested (61%) could complement a lethal growth defect and replace their yeast orthologs with minimal effects on growth rate. Replaceability was determined on a pathway-by-pathway basis; codon usage, abundance, and sequence similarity contributed predictive power. The heme biosynthesis pathway was particularly amenable to inter-kingdom exchange, with each yeast enzyme replaceable by its bacterial, human, or plant ortholog, suggesting it as a near-universally swappable pathway.

2.2 Introduction

Despite over 2 billion years of divergence, eukaryotes and prokaryotes still share hundreds of genes[41, 35, 39, 40]. Though these ancient genes are identifiable as orthologs at the sequence level, the preservation of original protein function across such deep timescales has not been systematically explored. The function of certain genes could potentially become frozen in place in the course of evolution, sheltered from lineage-specific functional alterations introduced by mutations, gene fusions, and non-orthologous gene displacements. Such functionally frozen genes would in principle be able to substitute for their least-diverged ortholog in any other species. Searching for such gene replaceability between species thus serves to test a core assumption of the ortholog-function conjecture: that orthologs retain ancestral function[42]. This conjecture forms the basis of most modern biomedical research and is widely used to predict new gene function across organisms[43].

There are many individual examples of genes from one species functioning for their orthologous counterparts in a different species[44, 45], but this trend has only recently begun to be explored systematically, with several large-scale studies substituting human genes for yeast genes and confirming that many human orthologs can successfully replace their yeast counterparts[8, 34, 33]. At the level of evolutionary divergence of yeast and humans, such data demonstrate widespread functional conservation, even after 1 billion years of divergence. The ability of human genes to functionally replace their yeast orthologs is not strongly predicted by the similarity of sequences, but rather at

the level of specific pathways or processes, wherein all genes in a process or pathway tend to be similarly replaceable, or not[8].

However, in the timescale of evolution, yeast and humans are relatively similar – both eukaryotes that share thousands of genes and the majority of their core biological processes. Data on eukaryote – prokaryote functional gene replacement are sparse[45]. These cross-domain replacements represent a maximum test of the ability of genes to retain their ancestral function across time. Eukaryotic and bacterial genes have been, for the most part, evolving independently since at least the archaeal ancestor of eukaryotes endosymbiotically acquired its bacterial mitochondrion. In eukaryotes, the function of these genes would have had to survive the development of vastly different genome structures, cell division modalities, cell wall compositions, and subcellular compartmentalizations which occurred during eukaryogenesis. Prokaryotic and eukaryotic orthologs also diverged significantly at the amino acid sequence level[35] and evolved distinct expression patterns and codon usages[47, 46]. Nonetheless, eukaryotes and bacteria are known to use many of the same orthologs to perform the same metabolic enzymatic reactions[48, 49].

Thus, in order to more systematically determine the replaceability of orthologs across such deep timescales, we asked in this study how many conserved *E. coli* genes can successfully substitute for their yeast orthologs. We focused on those genes that are essential for viability in yeast, allowing us to assay for the complementation of otherwise lethal growth defects. We analysed many features of the proteins and ortholog pairs to identify which properties

best explained replaceability, finding that replaceability was often determined at the level of specific pathways and processes, with all genes in a pathway or process similarly replaceable. Start codon choice and eukaryote-specific subcellular localization were also critical determinants of replaceability. We discovered that certain core biological processes have remained largely unchanged since the last common ancestor of bacteria, yeast, and humans. In particular, heme biosynthesis pathway enzymes appear to be generally exchangeable between prokaryotic and eukaryotic organisms, broadly retaining ancestral functions across the tree of life over 2 billion years of independent evolution, even when accompanied by evolved changes in enzyme subcellular localization.

2.3 Results and discussion

2.3.1 Many *E. coli* genes successfully complement lethal defects in their yeast orthologs

Many *E. coli* genes successfully complement lethal defects in their yeast orthologs We focused our efforts on the set of genes with 1:1 orthology between *E. coli* and yeast and that are known to be essential for yeast growth in standard laboratory conditions (Figure 2.1A). Each *E. coli* open reading frame (ORF) was cloned into a single-copy yeast centromeric (CEN) plasmid under the transcriptional control of a constitutive GPD promoter. Complementation assays were carried out using two types of conditionally essential yeast alleles, consisting of temperature-sensitive (TS) haploid and heterozygous diploid deletion strains. In the case of the heterozygous diploid deletion strains, the respective yeast gene null allele could be genetically segregated via sporulation, allowing selection for haploid yeast with the null allele (selected for in the presence of the antibiotic G418) or the wild-type yeast gene (in the absence of G418) (Figure 2.1B, top panel). In the case of TS haploid yeast strains, the temperature sensitive yeast proteins functioned normally at the permissive temperature (25°C) but could be conditionally inactivated at the non-permissive temperature (36°C) in order to test for gene replaceability (Figure 2.1B, bottom panel). Overall, we could perform informative complementation assays for 51 of the 58 orthologs, as shown for the examples in Figure 2.1B.

Of the 51 *E. coli* genes tested, 25 successfully complemented lethal

growth defects in the corresponding yeast strains (Figure 2.2A, B and C; Supplementary file 1). In nearly all cases, despite plasmid-based expression of the complementing genes, the bacterialized strains grew comparably to the parental, wild type yeast strain, in both synthetic defined medium (SD-Ura+G418) (Figure 2.1C) and rich medium (YPD+G418a) (Figure 2.2D). We further verified complementation specificity by testing for plasmid loss (see Materials and methods and supplementary file 1) and sequence verifying all clones. We have previously demonstrated that plasmid-borne copies of yeast genes complemented their corresponding heterozygous diploid deletion alleles at a high rate (100% for 29 strains tested in [8]), but as an additional control, we repeated this test for six yeast strains where the *E. coli* gene failed to rescue, confirming that the corresponding yeast genes were able to complement the growth defect when expressed on a CEN plasmid under the control of the constitutive GPD promoter (Figure 2.3 and Sc-HEM1 as reported in Figure 2.8).

2.3.2 Mitochondrial localization and start codon choice both affect replaceability

Many eukaryotic orthologs of prokaryotic genes function in specific subcellular compartments absent from prokaryotes, and consistent with this trend, 15 of the 51 tested *E. coli* genes have mitochondrially-localized yeast orthologs[44]. Because all but one of these 15 genes were unable to replace their yeast ortholog, we reasoned that lack of mitochondria targeting might account for their failed complementation. We added the mitochondrial localization sig-

nal (MLS) from the yeast MIP1 gene to each of the 14 non-replaceable *E. coli* genes and repeated the complementation assays. Four genes could now functionally replace their yeast equivalents (Figures 2.4A, 2.5), restoring growth rates to be nearly or fully comparable with the parental strain (Figure 2.4B). We verified mitochondrial localization by fusing the *E. coli* Ec-MLS-HscB and Ec-MLS-IlvD proteins with enhanced green fluorescent protein (EGFP) and confirming correct trafficking of the EGFP-tagged proteins to yeast mitochondria (Figure 2.4C).

Bacterial genes also occasionally lack a standard ATG start codon, with ∼14% of all *E. coli* ORFs employing an alternative start codon[50]. Three of the tested non-replaceable *E. coli* genes used a GTG start codon while one used ATT. We therefore used site-directed mutagenesis to introduce canonical ATG start codons, then re-assayed for complementation. After changing their start codons to ATG, two of these four *E. coli* genes, Ec-rcsC and Ec-tadA, could now replace their yeast orthologs (Figure 2.4B).

Overall, after accounting for mitochondrial localization and alternative start codons and combining results from all assays, a total of 31 out of 51 tested *E. coli* genes could successfully replace their essential yeast orthologs (Figure 2.4). Thus, in a majority (61%) of our tests, both the current day prokaryotic and eukaryotic proteins must have retained their critical ancestral functions such that the prokaryotic proteins could carry out the essential roles of their eukaryotic orthologs well enough to support yeast cell growth. In one-fifth of the cases, replaceability depended on proper subcellular local-

ization or start codon choice to express the prokaryotic gene in the proper eukaryotic context.

2.3.3 Replaceability varies strongly across different biological processes

Given that we observed both replaceable and non-replaceable genes, we sought to determine properties of the tested genes that best explained successful replacements. We considered 22 features of the tested genes, including protein lengths, interactions, sequence similarities, codon usages, and expression levels. We calculated the predictive utility of each feature as the area under a Receiver Operating Characteristic curve (AUC) (Figure 2.6A; supplementary file 2). Notably, the extent of protein sequence similarity between orthologs was not a highly predictive feature. A large portion of the tested *E. coli* and yeast orthologs showed only 20-30% identical amino acid sequences and roughly half of these genes were replaceable; in contrast, the three most divergent orthologs replaced, each showing less than 20% identity (Figure 2.6B). As we observed a non-monotonic relationship between sequence identity and replaceability, potentially explained by replaceability differences among different functional categories of genes, we tested for the enrichment of particular GO Biological Process (defined by Gene Ontology Slim annotations[51] or KEGG categories[52] within the individual bins of sequence identity in Figure 2.6B. Aside from an enrichment in glucose metabolism genes (3 of the 7) in the 40-50% identity range, we did not find evidence for strong pathway-specific biases that would explain the observed relationship between sequence identity

and replaceability. We did observe moderate predictive power for some measures of codon bias, especially those related to codon optimality within *E. coli*, and less so for codon optimality within a yeast context; more highly optimized *E. coli* codon usage correlated with a lower replaceability rate.

Instead, the strongest predictive features related to specific pathways and processes, much as we and others have observed for successful humanization of yeast[8, 34, 33]. This trend was most evident in the observation that a gene was more likely to replace (or not) if it had a higher fraction of interaction partners that also replaced (or not). Consequently, different biological processes (as defined by GO) displayed varied replaceability, with metabolic processes being largely replaceable, while processes known to be divergent, including ribosomal processing, were much less replaceable (Figure 2.6C). This trend suggests an explanation for why optimized *E. coli* codons predicted worse replaceability, as *E. coli* genes with optimized codons predominantly tend to be highly expressed ribosomal and translational proteins[53]. This is thus consistent with the notion that replaceability is determined at the level of the pathway or process, with codon choice and gene expression levels reflecting functional constraints of that process. Combining all of these features into a single predictor (after accounting for mitochondrial localization and alternative start codons), using a random forest classifier, improved our predictive power to a 0.79 AUC (Figure 2.6A), demonstrating that the features we investigated provide moderately orthogonal predictive information.

2.3.4 Each yeast heme biosynthesis enzyme can be replaced by its *E. coli* equivalent, irrespective of orthology or localization

Nearly all the genes that we tested from the heme biosynthesis pathway were replaceable by their *E. coli* orthologs, which in combination with the evidence that replaceability was determined at the level of processes, led us to investigate the heme pathway in more depth. Most of the enzymatic reactions in the heme biosynthesis pathway are identical between *E. coli* and yeast, but there are clear differences in the way this pathway functions between the species[54]. First, heme biosynthesis pathway precursors differ: Yeast condense succinyl-CoA and glycine to produce delta-aminolevulinate in a single enzymatic step catalyzed by Sc-Hem1, while *E. coli* produces delta-aminolevulinate in two steps using glutamyl-tRNA as a precursor[55]. Second, the bacterial heme pathway is largely cytosolic but in yeast it is partitioned between the mitochondria and cytosol (Figure 2.7A). We thus next considered these two key pathway differences in more detail. As a control, we first expressed the corresponding yeast genes on plasmids either under the control of constitutive GPD or the native yeast promoter[56] to test the effect of constitutive expression on functional replaceability. Except for Sc-HEM4, which showed toxicity when expressed constitutively, all the other yeast genes showed functional replaceability irrespective of the mode of expression (Figure 2.8).

In our initial screen, the *E. coli* ortholog of Sc-HEM1, Ec-kbL, failed to replace the yeast gene, an observation consistent with prior data showing that Ec-kbL does not take part in *E. coli* heme biosynthesis, but rather carries

out an unrelated but mechanistically-similar oxido-reductase reaction involved in L-threonine degradation ([58, 57]). Instead, a two-step enzymatic reaction by *E. coli* proteins Ec-HemA and Ec-HemL produces the heme precursor, delta-aminolevulinate ([60, 59]). Since the initial steps of the pathway are localized to the mitochondria, we added the Sc-MIP1 MLS to the 5' ends of these genes and expressed them simultaneously in the Sc-HEM1 heterozygous diploid deletion strain. Co-expression of the two *E. coli* genes successfully replaced yeast gene function (Figure 2.9A). Additionally, two enzymes, Ec-HemD and Ec-HemG, were not identified as orthologs between *E. coli* and yeast, despite carrying out identical reactions to Sc-Hem4 and Sc-Hem14, respectively. Expression of these non-orthologous but functionally analogous *E. coli* genes in the respective yeast deletion strains showed that they were indeed able to successfully replace the yeast genes (Figure 2.9B). For these enzymes, the key determinants for successful replacement are thus their enzymatic reactions, rather than any other aspects of the genes.

Sc-Hem14 and Sc-Hem15 carry out the final two steps in yeast heme biosynthesis and are localized to the mitochondria ([44, 61]) (Figure 2.7A). Both genes were replaceable by the *E. coli* genes carrying out the analogous reactions, Ec-HemG (Figure 2.9B) and Ec-HemH (Figure 2.1C), despite the lack of targeting sequences for mitochondrial localization. As *E. coli* lack mitochondria, and Ec-HemG and Ec-HemH are both predicted to localize to the plasma membrane in *E. coli*[62], we thus assayed their localization in yeast when expressed as EGFP-fusion proteins. Strikingly, both localized to the

yeast plasma membrane (Figure 2.10). In spite of failing to localize to the yeast mitochondria, the bacterialized strains grew well compared to wild type yeast (Figure 2.10), suggesting that mitochondrial localization is not an absolute requirement for their functions, as many heme pathway intermediates are cytosolic. However, concurrent bacterialization of both yeast genes resulted in a viable but defective yeast strain (Figure 2.9C), suggesting that the fitness cost of mis-localizing both proteins is not tolerated well, potentially due to cumulative effects of reduced efficiency of the bacterial proteins, altered allosteric regulation in yeast, or the accumulation of heme precursors in the wrong compartment (cytosol)[55].

Because heterologous expression using a constitutive promoter could be compensating for more subtle functional differences, we also wished to measure complementation after placing the bacterial orthologs under control of the native yeast gene regulation. We thus used CRISPR/Cas9-based precision genome engineering to genomically replace each of the heme biosynthesis pathway genes in turn in yeast (except Sc-HEM12) with its respective *E. coli* counterpart, from start to stop codon, while retaining the native promoters, terminators, and chromosomal context of the yeast genes (Figure 2.7B, 2.11). All strains but two grew comparably to the wild-type; the Sc-hem14Δ::Ec-hemG and Sc-hem15Δ::Ec-hemH strains showed modest growth defects (Figure 2.7B). Because these two yeast proteins are known to be mitochondrially localized[44], we re-engineered each of the Ec-hemG and Ec-hemH ORFs into the yeast chromosome such

that each gene's native yeast MLS was retained (Sc-hem14Δ::Ec-MLS-hemG and Sc-hem15Δ::Ec-MLS-hemH). The addition of the yeast MLS to each *E. coli* ORF completely ameliorated growth defects from the ORFs alone (Figure 2.7B).

Thus, the yeast heme biosynthesis pathway appears entirely replaceable, one gene at a time, by their corresponding bacterial genes, whether expressed constitutively from plasmids or directly integrated into chromosomes under native yeast transcriptional regulation. The extent of replaceability strongly suggests that ancestral functions in these genes (with the obvious exception of the non-orthologous steps) have remained intact and unaltered, at least in terms of critical, enzymatic functionality. Mitochondrial localization of several of the enzymes, while needed to fully recover growth rates, is not essential for viability.

2.3.5 Bacterialization with the *E. coli* ferrochelatase induces a yeast phenotype resembling human porphyria

Ec-hemH and Sc-HEM15 encode ferrochelatase, the enzyme responsible for adding iron to the porphyrin ring of protoporphyrin IX to produce protoheme (Figure 2.7A). In the course of constructing the CRISPR-edited yeast strains, we noticed that the Sc-hem15Δ::Ec-hemH yeast strain turned pink on a standard YPD agar medium upon prolonged incubation of 3–4 days (Figure 2.12A). This phenotype was consistent across all independently obtained, sequence verified yeast clones. The pink phenotype decreased dramatically

in the Sc-hem15Δ::Ec-MLS-hemH strains in which Ec-HemH was correctly localized to the mitochondria by addition of an MLS.

We speculated that the pink phenotype was likely due to aberrant accumulation of porphyrin intermediates, presumably leading to their secretion, as we observed that the pigment could be washed off the cells. Therefore, we chemically extracted the pink pigment from Sc-hem15Δ::Ec-hemH,0 Sc-hem15Δ::Ec-MLS-hemH and wild type yeast cells (Materials and methods) and performed fluorescence spectroscopy to determine that the pigment likely corresponds to protoporphyrin IX (Figures 2.12B, 2.13).

In order to determine whether protein mis-localization contributed to the phenotype, we removed the MLS from the native yeast gene. Several clones of the Sc-ΔMLS-HEM15 yeast strain displayed similar extracellular pigment (Figures 2.12B, 2.13). These results suggest that mislocalized plasma membrane-bound Ec-HemH in yeast does not convert protoporphyrin IX to protoheme efficiently, resulting in the accumulation and secretion of protoporphyrin IX. We further tested this line of reasoning by deleting the gene for the preceding step in the pathway, Sc-HEM14, which encodes the enzyme protoporphyrinogen oxidase and is responsible for making protoporphyrin IX. Using CRISPR, we deleted the Sc-HEM14 ORF in wild type BY4741, Sc-hem15Δ::Ec-HemH, and Sc-hem15Δ::Ec-MLS-HemH strains. Consistent with protoporphyrin IX being the pink pigment in the Sc-hem15Δ::Ec-HemH strain, the Sc-hem15Δ::Ec-HemH hem14Δ strain lost the pink phenotype, even after growing for 6 days. Moreover, we observed that all strains carrying

the hem14Δ allele were in fact significantly paler than even wild type BY4741 cells, presumably reflecting extensive protoporphyrin IX depletion in these cells (Figure 2.14).

In humans, disrupting heme biosynthesis leads to the disease porphyria, and the secretion of porphyrin intermediates is specifically observed in a subtype known as protoporphyria[63], wherein reduced activity of the human heme pathway protein Hs-FECH leads to accumulation and subsequent secretion of protoporphyrin IX into surrounding tissues. Our data suggest that yeast protein localization and protoporphyrin secretion phenotypes might in the future be exploited to investigate disease-causing mutations in human Hs-FECH, even in cases where disease variants do not show any discernible growth defect in yeast.

2.3.6 Most yeast heme biosynthesis enzymes can also be successfully plant-ized

The data above show that genes in the yeast heme biosynthesis pathway can be replaced by their bacterial counterparts, extending earlier studies demonstrating that some heme biosynthesis genes can also be humanized ([8, 34, 64]). Given the ancient conservation of this pathway, we sought to further expand our investigation of its replaceability by swapping the corresponding genes from the plant Arabidopsis thaliana into yeast. In plants, heme biosynthesis enzymes form precursors for chlorophyll, and the pathway is largely chloroplast-localized, in contrast to compartmentalization of the heme biosyn-

thetic pathway between the mitochondria and cytosol in many other eukaryotes ([58, 51, 65]). Nonetheless, the fact that Arabidopsis ferrochelatase was cloned by complementing a mutant yeast phenotype suggests that other heme pathway genes might also successfully replace the yeast genes[66].

The first enzymatic step in the plant heme biosynthetic pathway is similar to bacteria, a two-step reaction using glutamyl-tRNA as a substrate (Figure 2.15A,B)[67]. We expressed both plant genes, At-HEMA1 and At-GSA2, simultaneously and were able to functionally replace the corresponding yeast gene function. Neither protein, when individually expressed, could functionally replace the yeast gene (2.16A).

In Arabidopsis, unlike for the case of *E. coli*, a majority of genes in the heme biosynthesis pathway have acquired lineage-specific amplifications, resulting in two co-orthologs for each single yeast gene (Figure 2.15B). In these cases, we tested both co-orthologs individually for replaceability; all replaced successfully, with the exception of one case where only one replaced (At-CPX1 replaced while At-CPX2 did not), and one case where neither replaced (At-HEME1 and At-HEME2) (Figures 2.15B and 2.16B).

Because the plant heme biosynthesis pathway builds precursors for chlorophyll synthesis[68, 69], this pathway, especially the penultimate step producing protoporphyrin IX, is the target of many commercial herbicides. Both Arabidopsis paralogs that we tested, At-PPOX1 and At-PPOX2, could efficiently complement the yeast gene responsible for this critical step, Sc-HEM14 (Figure 2.16B). To confirm the generality of these results, we further tested the

soybean (Glycine max) ortholog Gm-HEMG in yeast. As for each of the Arabidopsis paralogs, the single soybean ortholog also successfully complemented the Sc-HEM14 deletion growth defect (Figure 2.16B).

It is noteworthy that plant heme biosynthesis genes harbor chloroplast localization sequences[58], and we did not remove these for our complementation experiments. We speculated that the chloroplast leader peptides might be recognized and localized by the yeast mitochondrial localization machinery, so we constructed EGFP-fusions of the plant enzymes and assayed their localization by fluorescence microscopy. EGFP fusions of At-PPOX1 and At-FC1 showed clear mitochondrial localization in yeast (Figure 2.17A). At-FC1 additionally showed amorphous aggregates in some yeast cells, suggesting localization might occasionally be imperfect. Nonetheless, both EGFP-tagged genes were able to efficiently rescue the growth defect of the corresponding yeast gene deletion (Figure 2.17A). Thus, these plant chloroplast localization signals appear to be recognized and processed as mitochondrial localization signals in yeast.

These findings suggested that plant versions of cytosolic yeast heme biosynthesis pathway proteins could potentially be mis-localizing to the mitochondria in yeast (Figure 2.7A). Indeed, At-HEMC only weakly replaced the yeast gene, Sc-HEM3. We found that removing the chloroplast localization signal (CLS) from At-HEMC markedly enhanced its ability to functionally replace its yeast ortholog (Figure 2.16B). In contrast, neither of two Arabidopsis paralogs, At-HEME1 and At-HEME2, could functionally replace their yeast

ortholog, Sc-HEM12, even after removing their CLS sequences, or even when co-expressed in the yeast strain (Figure 2.16B). We speculate that there could be several other reasons why complementation failed, including unknown intermediate reactions, required localization in a special compartment (e.g. chloroplast) or different transcriptional/translational regulation in plants that might contribute to the lack of functional replaceability.

2.3.7 Each yeast heme biosynthesis enzyme can be replaced by its human ortholog

Earlier studies have shown successful replacement of the yeast heme biosynthesis genes by their human orthologs Hs-ALAD ([64]), Hs-HMBS, Hs-CPOX and Hs-FECH[8], while Hs-UROS expression resulted in toxicity and Hs-UROD failed to replace its yeast ortholog[8, 34]. We, therefore, sought to complete tests of the remaining human genes in the pathway. In the case of Hs-UROS, we reasoned that toxicity was due to expression from the heterologous constitutive promoter (2.18A). Indeed, similar to the results obtained with the yeast version of this gene (Figure 2.8, Sc-HEM4), we found that toxicity could be abrogated by inserting the human gene at the native yeast chromosomal locus, thus providing native yeast gene expression and regulation for the human ORF (2.18B). This suggests that, at least in yeast, this step is regulated transcriptionally for optimal function. We also found that the human ORFeome clone of Hs-UROD contained a mutation (G303V) that when reverted to wild-type sequence allowed it to replace the yeast gene (Figure 2.18C,D), and we additionally confirmed that human Hs-PPOX could

complement the severe growth defect of the yeast Sc-HEM14 deletion strain (2.15C, 2.18D). Finally, in humans, the initial step of heme biosynthesis is identical to that of yeast (Sc-HEM1) but is encoded by two co-orthologs, Hs-ALAS1 and Hs-ALAS2. We found that both of these human genes could individually replace the yeast gene function (2.15C, 2.18D).

The subcellular localization of heme biosynthesis differs slightly between humans and yeast, such that the last three proteins in the human heme biosynthesis pathway are mitochondrially localized, as opposed to only the last two in yeast[71, 70]. As all three of these genes replaced, we tested if the human genes were localized to the mitochondria in yeast. Indeed, EGFP-tagged Hs-FECH, Hs-PPOX, and Hs-CPOX all localized to mitochondria in yeast (2.17B) and efficiently rescued the growth defect of the corresponding yeast gene deletion (2.17B), confirming that the human mitochondrial localization signal is recognizable by the yeast localization machinery. Thus, across our attempts to humanize, plantize, and bacterialize this pathway, the presence of mitochondrial leader peptides from the human genes and the chloroplast leader peptides from the plant genes, as well as the absence of bacterial leaders, all overrode the native yeast localization of the heme biosynthesis pathway. However, the pathway function was largely resilient to these effects, with the exception of protoporphyrin IX accumulation in the mislocalized bacterialized strains (2.12).

2.3.8 Heme biosynthesis is a near-universally swappable pathway

As illustrated in 2.19, the heme pathway has had a complicated evolutionary trajectory in eukaryotes due to endosymbiotic events, which has served to increase its similarity between bacteria and eukaryotes[72]. During eukaryogenesis, early eukaryotes adopted a large portion of the bacteria-like heme biosynthesis pathway of their endosymbiont mitochondria. The subsequent endosymbiotic acquisition of chloroplasts along the plant lineage[73] resulted in redundancy between mitochondrial-origin and chloroplast-origin portions of their heme biosynthesis pathways, a state that can be observed today in Euglena, a non-plant, photosynthetic eukaryote with more recently acquired chloroplasts[72]. Over time, plants kept the chloroplastic system and lost most of the mitochondrial system. These evolutionary transfers may have been possible due the apparent modularity of the heme pathway, which we observe in its high tolerance for substituting genes or enzymatic functions across species.

Our data demonstrate that despite 2 billion years of divergence from their last common ancestor, heme biosynthesis genes are still carrying out a conserved and necessary function that can be swapped into yeast with minimal effect on growth and irrespective of orthology and subcellular localization. Taking these data together with literature studies showing successful replacement of the E. coli Ec-hemG gene by the plant or human Hs-PPOX gene[75, 74, 76], and that introducing the protoporphyrinogen oxidase from *Bacillus subtilis* into plants improves yields[77], heme biosynthesis thus appears to be

a pathway whose genes are freely exchangeable between bacteria, plants (with the exception of At-HEME), humans, and yeast2.19.

2.3.9 Conclusions

In conclusion, in order to discern whether orthology strictly confers function across deep evolutionary distances, we systematically tested *E. coli* genes with 1:1 orthology to essential yeast genes for their ability to functionally replace their yeast counterparts. We discovered that ∼61% (31/51) of the tested *E. coli* and yeast genes still retain ancestral function to a sufficient extent that the bacterial genes efficiently replace their yeast equivalents. Eukaryote-specific features such as subcellular localization (4 of 14) and proper start codon usage (2 of 4) were critical for swappability for some of the *E. coli* orthologs. Our analysis of replaceable/non-replaceable orthologous pairs revealed that amino acid sequence similarity was not the most important property, consistent with a general trend for sequence conservation to often more strongly reflect other attributes of protein function (e.g., abundance and protein-specific functional constraints) ([78, 79]). Rather, the top predictors of replaceability were features attributed to specific gene modules. These results largely agree with previously published work on humanization of yeast genes[8, 33, 34], suggesting that functional replaceability is predominantly determined at the level of pathways and processes, even across very large evolutionary distances. As our assays can be considered a form of forced horizontal gene transfer, our results provide support for the "complexity hy-

pothesis"[80], which posits that informational (transcription, translation, etc.) genes are less likely to be horizontally transferred than those genes that are operational (metabolism, housekeeping, etc.). Consistent with this expectation, we see metabolism-associated genes replacing more often than those involved in "informational" processes like transcription or translation.

In the course of these studies, we found that heme biosynthetic reactions were entirely replaceable across the prokaryote-eukaryote divide, despite non-orthologous functional displacement and lack of eukaryotic subcellular localization by native *E. coli* genes (2.19). Although the archaeal pathway is considerably diverged, our studies across bacteria and eukaryotes showed a high degree of replaceability: Plant heme biosynthesis enzymes functionally replaced yeast enzymes in all but one reaction. Swaps of the corresponding human enzymes into yeast in this and prior studies all suggest that heme biosynthesis is a near universally replaceable pathway.

Our results thus demonstrate that orthologous genes carry out similar functions that allow for their ability to functionally replace each other across even the 2 billion year evolutionary rift separating prokaryotes and eukaryotes from their last common ancestor. These swaps allow engineering of orthologous pathways in model organisms highly amenable to genetic perturbations, like yeast and bacteria, for further characterization.

2.4 Materials and methods

2.4.1 Construction of ORFs from bacteria, plants, and humans in yeast expression vectors

Refer to Supplementary file 3 for all the primers used in this study.

2.4.1.1 *E.coli* ORF yeast expression vectors

Initial *E. coli* ORF primers were designed such that the 3' ends of the primers had homology to *E. coli* genes and 5' ends contained a universal flanking sequence. A second round of PCR was performed with primers recognizing the universal flanking sequence and also having 5' ends corresponding to gateway compatible attL1 (or attB1) and attL2 (or attB2) sequences on the forward and reverse primers, respectively. Resulting PCR products from attL sequence containing primers were directly cloned via gateway LR cloning (ThermoFisher Scientific) into yeast destination vector pAG416GPD-ccdB (Addgene) to create expression clones. PCR products from attB primers were subcloned via gateway BP cloning into vector pDONR221 (ThermoFisher Scientific) to create entry clones. These entry clones were then cloned via gateway LR to the pAG416GPD-ccdB destination vector to create expression clones. Some *E. coli* genes were synthesized as gBlocks from IDT and made gateway compatible by adding attL1 and attL2 sequences at the 5' and 3' ends, respectively, making them compatible for direct LR cloning to create expression vectors.

2.4.1.2 Plant ORF yeast expression vectors

Arabidopsis thaliana ORFs were PCR amplified from cDNA obtained as a kind gift from Dr. Jeffrey Chen (UT Austin), using primers specific to each gene and containing gateway compatible attL1 and attL2 sequences at the 5' and 3' ends respectively (Supplementary file 3). PCR products were directly cloned into the yeast expression vector pAG416GPD-ccdB by LR gateway cloning (using LR clonase II from Invitrogen). At-HEME1 and At-HEMB2 were synthesized as gBlocks from Integrated DNA Technologies (IDT).

2.4.1.3 Plant ORF yeast expression vectors without the chloroplast localization signal

In order to remove the chloroplast localization signal from the plant proteins At-HEMC, At-HEME1 and At-HEME2, we first performed amino acid sequence alignment with the bacterial and yeast orthologs to identify unaligned N-terminal sequence. We attributed the non-alignment to the presence of chloroplast localization signal (CLS) sequence. We also used the TargetP 1.1 signal peptide predictor[81] to corroborate the sequence alignments. In the case of At-HEMC, 68 N-terminal amino acids were deleted while retaining ATG start codon. Similarly, in the case of At-HEME1 and At-HEME2, 47 N-terminal amino acids were deleted while retaining the ATG start codon. We synthesized these genes as gBlocks (IDT) with attB1 and attB2 attachment sites flanking their 5' and 3' ends, respectively, then subcloned the gBlocks into the entry clone pDONR221, sequence-verified the clones, and LR cloned

the genes into yeast expression vector pAG416GPD-ccdB.

2.4.1.4 Plant ORF yeast expression vectors for co-expression of At-HEME1 and At-HEME2

At-HEME1 and At-HEME2 were cloned (with or without CLS) into the destination vectors pAG416GPD-ccdB and pCMY41 (kind gift of Christopher Yellman; pCMY41 is identical to pAG416GPD-ccdB but carries a hygromycin-resistance cassette), allowing us to co-transform two plasmids and select for the double plasmid transformants on synthetic defined medium, -Ura+Hyg (200 µg/ml).

2.4.1.5 Human ORF yeast expression vectors

Human ORF's were obtained from the ORFeome collection (GE Dharmacon) and sequenced to verify correct, full-length clones. In the case of human Hs-UROD, the ORFeome clone contained a loss-of-function mutation (G303V), so wild-type human Hs-UROD was synthesized as a gBlock fragment (IDT) and used as a PCR template, amplifying the gene using primers with flanking gateway compatible sites attL1 and attL2 at the 5' and 3' ends respectively (Supplementary file 3). The PCR product was subcloned by LR reaction into the yeast expression vector pAG416GPD-ccdB.

2.4.1.6 Yeast ORF yeast expression vectors

Yeast ORFs were amplified using PCR from genomic DNA of yeast strain BY4741, and gateway compatible attL1 and attL2 sequences added

by PCR to the amplicons 5' and 3' ends, respectively (Supplementary file 3). The resulting PCR products were subcloned by LR reaction into the yeast expression vector pAG416GPD-ccdB. Several yeast heme biosynthesis genes were first cloned in pENTR/SD/D-TOPO plasmid (Invitrogen) to obtain gateway entry clones (refer to Supplementary file 3 for primers). These clones were sequence-verified and then used to generate yeast expression vectors by LR reaction into the vector pAG416GPD-ccdB.

2.4.1.7 Mitochondrially-localized *E. coli* ORF yeast expression vectors

We added the MLS from the yeast MIP1 gene to the 5' end of selected *E. coli* ORFs via PCR, using an ORF-specific ultramer containing the full MLS-coding sequence at the 5' end such that MLS was in frame with the coding sequence of the *E. coli* gene while removing the *E. coli* gene start codon (Supplementary file 3). Each PCR product was then used as a template to add gateway cloning attachment sites attL1 and attL2, followed by LR gateway cloning into pAG416GPD-ccdB to generate yeast expression vectors.

2.4.1.8 EGFP tagged *E. coli*/plant/human ORF yeast expression vectors

Using PCR, we amplified *E. coli*/plant/human ORFs without their respective stop-codons while also adding attB1 and attB2 gateway attachment sites at the 5' and 3' ends of each PCR product (Supplementary file 3). The resulting PCR fragments were subcloned into plasmid pDONR221

to generate gateway entry clones using the BP gateway cloning reaction. Each entry clone was subjected to the LR cloning reaction in order to generate a carboxy-terminal EGFP-tagged yeast expression clones in the pAG416GPD-ccdB-EGFP destination vector.

2.4.1.9 Converting *E. coli* ORF yeast expression vectors with alternative start codons to ATG start codon

We introduced ATG start codons by PCR mutagenesis, employing ATG-containing primers (Supplementary file 3) to amplify and simultaneously add gateway cloning attachment sites attL1 and attL2 to the 5' and 3' ends of the PCR products, respectively, then subcloning these products by the LR gateway cloning reaction into the pAG416GPD-ccdB plasmid in order to construct yeast expression vectors.

2.4.1.10 *E.coli* and Arabidopsis two-gene expression vectors for complementing a yeast Sc-HEM1 deletion

E. coli genes Ec-hemA, Ec-hemL and plant genes At-HEMA1, At-GSA2 were PCR amplified from genomic DNA (E. coli) or gBlocks obtained from IDT (Arabidopsis). For *E. coli* genes, we also added an MLS at the 5' end of the PCR products. These PCRs were made Golden Gate compatible by introducing Bsmb1 sites and cloned individually in pYTK001 (Supplementary file 3). In the case of At-HEMA1, the gBlock was synthesized to mutate an internal BsmBI site such that it doesn't affect the protein sequence. Clones were sequence verified prior to assembly ([23]). Individual

transcription units for each of the genes were obtained by Golden Gate assembly using the pYTK001-entry clone containing the E. coli or plant gene, along with pYTK vectorscontributing promoters and terminators. In the case of Ec-hemA and At-HEMA1 transcription units (TU1's), the pHHF2 promoter was contributed by pYTK012 and tADH1 terminator by pYTK053. In the case of Ec-hemL and At-GSA2 transcription units (TU2's), the pTEF1 promoter was contributed by pYTK013 and tSSA1 terminator by pYTK052. Unique contigs for directional assembly were obtained from pYTK002 (ConLS) and pYTK067 (ConR1) for TU1. For TU2, the unique contigs were obtained from pYTK003 (ConL1) and pYTK072 (ConRE). The individual transcription units (TU1 and TU2) were then assembled in a single yeast CEN6-URA vector via Golden Gate assembly with BsmbI.

All clones were sequence-verified using the University of Texas Genomic Sequencing and Analysis Facility.

2.4.2 Functional complementation assays

Gene replaceability was tested using available yeast strains from two yeast strain collections, the temperature-sensitive (TS) collection ([82]) and the heterozygous diploid deletion magic marker collection ([83]), as follows:

2.4.2.1 Temperature-sensitive (TS) collection assays

Typically, yeast strains in this collection grow at permissive temperatures (22–26°C) but cannot grow at restrictive temperatures (35–37°C).

Growth at restrictive temperatures thus allows for the identification of foreign genes that complement the yeast defect. We tested for replaceability in temperature-sensitive yeast strains as follows:

The strains were transformed with either an empty vector control (pAG416GPD-ccdB) or with the clone expressing the foreign gene. The transformants were plated on:

1. Ura dextrose medium at the permissive temperature (25°C), serving as a control for transformation efficiency and/or toxicity since both the yeast and the human gene are expressed.
2. Ura dextrose medium at the non-permissive temperature (36°C), testing for functional replacement under conditions in which the corresponding yeast gene is non-functional.

2.4.2.2 Heterozygous diploid deletion magic marker collection assays

The yeast heterozygous diploid deletion magic marker collection comprises yeast strains that harbor a deletion of one copy of a yeast gene replaced with a KanMX cassette. The strains also carry a magic marker or synthetic genetic array (SGA) cassette at the can1 locus, which enables selection for haploid cells on magic marker (MM) medium (-His -Arg -Leu +Can) post-sporulation with or without antibiotic G418 (200 µg/ml). Haploid a-type spores that harbor a wild type gene grow normally on magic marker (MM) medium without G418 and provide a test of sporulation efficiency and toxic-

ity, if any, associated with heterologous expression of the foreign gene (using a -Ura selection marker in this study). Growth of haploid spores on MM medium in the presence of G418 selects for yeast cells that harbor the relevant gene deletion while testing for complementation by the foreign gene.

Expression clones or empty vector controls were transformed into appropriate strains and selected on -Ura G418 medium in a 96-well format. (Toxicity was inferred from a repeated failure to obtain transformants in the case of expression clones compared to the empty vector control) Transformants were re-plated on GNA-rich pre-sporulation medium containing G418 (200 µg/ml) and histidine (50 mg/l). Individual colonies were inoculated in liquid sporulation medium containing 0.1% potassium acetate, 0.005% Zinc acetate, and incubated with vigorous shaking at 25°C for 3–5 days, after which sporulation efficiency was estimated by microscopy, and the mixture re-suspended in water and equally plated on two assay conditions:

1. "G418 minus" magic marker dextrose medium (-His-Arg-Leu+Can-Ura), incubated at 30°C. The haploid spores that carry the wild-type yeast gene grow in this medium acting as a control for sporulation efficiency. This condition also assays for toxicity if the haploid spores carrying expression vectors fail to grow.
2. "G418 plus" magic marker dextrose medium (-His-Arg-Leu+Can-Ura) containing 200 µg/ml G418. The resulting haploid deletion strain is expected not to grow, providing an assay of replaceability for strains carrying the expression vector. Cases with approximately equal num-

bers of cells growing in the absence or presence of G418 were considered functional replacements.

Positive assays were verified independently. Individual colonies were isolated from selective plates and used for growth assays on YPD or magic marker medium with G418 (Figure 2.1B, 2.2A). After growth on YPD with G418, each strain was spotted on 5-FOA agar to test plasmid dependency (Supplementary file 1). Only one strain (Ec-valS) failed that test.

2.4.3 Ortholog inference

Genes with 1:1 orthology between yeast and *E. coli* were obtained from the Inparanoid 8 webserver ([84]) and filtered to an only yeast-essential set. Orthologs to these selected yeast genes in human and Arabidopsis were downloaded from Inparanoid 8 and further refined by comparison to orthology calculations by eggNOG4.5 ([85]), OMA ([86]), and reference to the evolutionary history of the heme pathway in photosynthetic organisms ([73]).

2.4.4 Computational analyses of replaceability

2.4.4.1 Feature assembly

- Sequence features. Protein sequence features were calculated using UniProt[58] proteomes from the respective species downloaded in March 2015. *E. coli* nucleotide sequence features were calculated using EcoGene ([87]) sequences downloaded April 2015.
  - The number of amino acids in the respective protein.

[Sc|Ec]_Length

- Calculated as the difference of the amino acid length of the *E. coli* protein subtracted from the length of the S. cerevisiae ortholog.

  Sc-Ec_LengthDifference

- Calculated as the absolute value of the above length difference.

  Sc-Ec_AbsLengthDifference

- The fraction of identical residues (PercentID) or similar residues (PercentSimilarity) in a global alignment (NWalign, `http://zhanglab.ccmb.med.umich.edu/NW-align/`) of the respective orthologs, as a function of the longest of the two (Longest) or the length of the aligned region (Aligned).

  Sc-Ec_PercentIDAligned

  Sc-Ec_PercentIDLongest

  Sc-Ec_PercentSimilarityAligned

  Sc-Ec_PercentSimilarityLongest

- The Codon Adaptation Index (CAI), Codon Bias Index (CBI), or Frequency of OPtimal codons (FOP) for the respective *E. coli* gene, calculated using the *E. coli* optimal codon table (Ec_) or S. cerevisiae optimal codon table (Ec_Sc) using codonw (`http://sourceforge.net/projects/codonw/`).

  Ec_CAI

  Ec_CBI

  Ec_FOP

Ec_ScCAI

Ec_ScCBI

Ec_ScFOP

- Abundance features.
  - Yeast (Sc) protein abundance data was taken from [88]. Yeast Transcript and RPF abundance were taken from [89]. Yeast RPF abundance is calculated as the ratio of RPF reads to Transcript reads for a given gene. *E. coli* data was taken from [90] (average iBAQ abundance only).

    Sc_TranscriptAbundance

    Sc_ProteinAbundance

    Sc_RPFAbundance

    Sc_TranslationEfficiency

    Ec_ProteinAbundance

- Network features.
  - Calculated from interactions present in BIOGRID 3.1.93 ([91]). 'BIOGRID' was calculated using only those interactions annotated as 'physical interactions', while 'BIOGRID-LT' was calculated using the subset of physical interactions found only by low-throughput experiments.

    Sc_BIOGRID-Betweenness

    Sc_BIOGRID-Clustering

    Sc_BIOGRID-Degree

Sc_BIOGRID-SumLLS

Sc_BIOGRID-LT-Degree

Sc_BIOGRID-LT-SumLLS

Sc_BIOGRID-LT-Betweenness

Sc_BIOGRID-LT-Clustering

- Calculated using the 'All Pathways' table from EcoCyc (https://ecocyc.org) downloaded in September 2016. To create the network, all pathways were considered 'cliques' so that all members of the pathway were annotated as interacting with all other members of the pathway. FractionComplementing is the fraction of interacting partners tested in our assays that were able to replace.

  Ec_EcoCyc_FractionComplementing

2.4.5 Calculating the predictive strength of features

The predictive power of each feature was calculated as the area under the receiver-operator characteristic curve (AUC) while treating each feature as an individual classifier. Each feature was sorted in both ascending and descending directions, retaining the direction providing an AUC>0.5. To assess significance, a shuffling procedure was performed as follows: For each feature, the replaceable/non-replaceable status of each ortholog pair was shuffled (retaining the original ratio of replaceable to non-replaceable assignments), and the AUC was calculated. The shuffling procedure was carried out 1000 times

for each feature, and the mean AUC values and their standard deviations are reported.

2.4.6 Combined classifier

A Random Forest classifier was constructed using all features and evaluated using 10-fold cross-validation. The random forest was constructed to have no maximum tree depth, and ties between similarly good attributes were broken randomly. The combined classifier was implemented using the Weka data-mining software ([92]).

2.4.7 Confocal microscopy

Yeast cultures expressing GFP-tagged bacterial, plant, or human genes were grown to an optical density (OD) of ∼1, then 500 l of the culture washed with 1X PBS, and mitochondria fluorescently labeled by adding 100 nM MitoTracker Red CMXRos (Invitrogen). The cells were incubated in the dark on a mildly shaking platform for 20 min at room temperature, then washed twice with 1X PBS and resuspended in 15 µl of 1X PBS for imaging by confocal microscopy, using a Zeiss LSM 710 confocal microscope with a Plan-Apochromat 63x/1.4 oil-immersion objective.

2.4.8 Quantitative growth curves

Yeast strains were either pre-cultured in liquid YPD or -Ura Dextrose selective medium for 2 hr or overnight respectively. The culture was diluted in

YPD or -Ura Dextrose medium to an OD of ∼0.1 in 100 or 150 l total volume in a 96-well plate. Plates were incubated in a Synergy H1 shaking incubating spectrophotometer (BioTek), measuring the optical density every 15 min over 48 hr. Growth curves were performed in triplicate for each strain by splitting the pre-culture into three independent cultures for each 48–60 hr time course.

2.4.9 Detection of heme pathway intermediate metabolites

Bacterialized Ec-hemH yeast strains were grown on YPD as lawns or large patches for 5 days (the phenotype manifests after several days of growth). Clumps of cells about 5–7 mm in diameter were collected with a toothpick and first suspended in water, then pelleted at 15,000 g for 30 s. This created a distinctive pale yellow yeast pellet, with the red pigment appearing in a small clump on top. The water was removed while carefully avoiding disruption of the red pigment pellet, after which we performed extractions with two different methods. The first method, based on [93], was to add 1 ml pyridine to each pellet, spinning down at 15,000 g for 30 s and recovering only the liquid fraction (cell debris would pellet down while the red pigment migrated into the liquid pyridine phase). The second referred to as "acetate extraction" in this text, was to extract with a 3:1 ethyl acetate:glacial acetic acid solution as described in [94].

We then measured the absorbance of the extractions in a transparent plastic 96-well plate on the (Synergy H1 from BioTek) on wavelengths from 223 nm to 998 nm, with 1 nm steps. We measured fluorescence on the same

instrument by exciting at 399 nm and measuring emission at 450 nm to 699 nm with 1 nm step. The spectra were compared with those shown in [95].

We also obtained protoporphyrin IX (Sigma-Aldrich, P8293-1G) and hemin B (Sigma-Aldrich, 51280–1G) and suspended these in acetate and pyridine to closely resemble the chemistry of our extractions. These solutions were measured alongside the extractions themselves as standards, in order to further confirm the identity of the molecules we detected.

2.4.10 Replacement of bacterial and human genes at their native yeast loci using CRISPR-Cas9

Genomic editing and replacement of yeast ORFs is described in greater detail at Bio-protocol[21].

2.4.10.1 Bacterializing yeast strains at native genomic loci using CRISPR

We inserted *E. coli* ORFs at their native yeast loci using CRISPR/Cas9-mediated double strand breaks (DSB) and homologous recombination. The integration was performed by chemically co-transforming yeast with a linear template DNA (Zymo Research - #T2001) and a plasmid carrying Cas9 and gRNA targeting the desired locus of integration (refer to Supplementary file 3). The transformed cells were plated on SD-Ura medium to select for successful transformation of the plasmid (CRISPR-induced DSBs act as partial selection against background), and screened for successful integration of the template via colony PCR using primers flanking the start

codon of the ORF (a forward primer annealing to the promoter and a reverse primer annealing to the *E. coli* ORF) (Figure 2.11).

The template DNA is a linear sequence containing the *E. coli* ORF, flanked by the yeast promoter and terminator which act as homology. In order to produce this template DNA, we designed primers for each gene that amplify the entire coding sequence of the *E. coli* ortholog, while also inserting flanking homologies to the yeast locus targeted. In most cases, we used primers 120 bp long, with about 20 bp shared with the *E. coli* gene and 100 bp of yeast homology. In cases where this template failed to integrate (such as Ec-hemC) we designed 200 bp primers with about 180 bp homology. For chimeric ORFs of *E. coli* genes Ec-hemG and Ec-hemH that retained the native yeast MLS, the template was produced by including the MLS in the forward primer sequence. We amplified the template DNA with PCR, purified it using the DNA Clean and Concentrator-25 kit (Zymo Research - #D4006); final elutions were done with water. We used 5 g DNA template per transformation, in cases where this failed we attempted it again with 10 g.

CRISPR plasmids were constructed using a Golden Gate-based cloning strategy as described in [23]. Briefly, for each yeast gene we designed two gRNA sequences using Geneious v9[96]; both sequences were selected from within the yeast ORF so as to exhibit high predicted efficiency with a low background activity for the rest of the yeast genome. We performed integration experiments separately for each gRNA, as often one of the gRNA sequences would have substantially lower efficiency than predicted. As per [23], each

gRNA sequence was first synthesized as an oligonucleotide (IDT), subcloned into intermediate plasmids, and eventually into a Cas9 plasmid carrying a Ura selectable marker, finally transforming 500 ng into yeast cells for the integration assay.

In order to construct yeast strain Sc-ΔMLS-HEM15, we started with Sc-hem15Δ::Ec-hemH yeast which had lost their CRISPR plasmid, and co-transformed them with CRISPR plasmids carrying gRNA that targets the Ec-hemH sequence, as well as template DNA created by amplifying the yeast Sc-HEM15 sequence from yeast genomic DNA. The MLS was deleted by designing template amplification primers which leave it out. This was necessary since the MLS sequence did not contain unique CRISPR targets, thus it was not possible to construct a CRISPR system that would cleave wild type Sc-HEM15 but not the desired Sc-ΔMLS-HEM15.

2.4.10.2 Humanizing Hs-UROS gene at the native yeast locus

We co-transformed the plasmid expressing Cas9 and gRNA targeting yeast Sc-HEM4 gene and repair PCR template that contains human Hs-UROS gene flanked by 100 bp of homologous sequence to the yeast Sc-HEM4 promoter and terminator region. The colonies that grew after the transformation of CRISPR plasmid and the repair template were verified for the human gene insertion using a forward primer outside the region of homology and reverse primer specific to the human gene. The positive PCR reaction with appropriate size (375 bp) confirmed the right clone.

2.4.11 Generation of Sc-HEM14 yeast deletion strains

Using CRISPR, we deleted the Sc-HEM14 ORF in wild type BY4741, Sc-hem15Δ::Ec-HemH, and Sc-hem15Δ::Ec-MLS-HemH strains. Specifically, we co-transformed the plasmid expressing Cas9 and gRNA targeting the yeast Sc-HEM14 gene with a 200 bp oligonucleotide repair template comprising 100 bp each of sequence matching the 5' and 3' UTRs of the Sc-HEM14 gene and selected for growth on SD-Ura medium. The resulting hem14Δ strains were confirmed by PCR using primers outside the region of homology. Supplementary file 3 provides relevant primers and oligos.

2.5 Supplementary material

2.5.1 Supplementary file 1

This file is a spreadsheet containing detailed results of complementation assays. It is available online at `https://dx.doi.org/10.7554/eLife.25093.022`.

2.5.2 Supplementary file 2

This file is a spreadsheet containing data used to calculate predictive features. It is available online at `https://dx.doi.org/10.7554/eLife.25093.023`.

2.5.3 Supplementary file 3

This file is a spreadsheet containing primers used in this study. It is available online at `https://dx.doi.org/10.7554/eLife.25093.024`.

2.6 Funding Information

This paper was supported by the following grants:

- National Institutes of Health R21 GM119021 to Edward M Marcotte.
- Cancer Prevention and Research Institute of Texas to Edward M Marcotte.
- Welch Foundation F1515 to Edward M Marcotte.
- National Institutes of Health R01 HD085901 to Edward M Marcotte.
- National Institutes of Health DP1 GM106408 to Edward M Marcotte.
- National Institutes of Health R01 DK110520 to Edward M Marcotte.
- National Institutes of Health R35 GM122480 to Edward M Marcotte.

2.7 Acknowledgements

This work was supported by grants from the NIH (R21 GM119021, R01 HD085901, DP1 GM106408, R01 DK110520, R35 GM122480), CPRIT, and the Welch foundation (F-1515) to EMM.

2.8 Additional information

2.8.1 Competing interests

The authors declare that no competing interests exist.

2.8.2 Author contributions

- AHK, Conceptualization, Data curation, Formal analysis, Supervision, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing.
- JML, Conceptualization, Data curation, Formal analysis, Supervision, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing.
- AA, Data curation, Validation, Investigation, Visualization.
- MS-J, Validation, Investigation, Visualization.
- CDM, Visualization, Methodology, Writing—review and editing.
- AZ, Visualization, Methodology.
- EMM, Conceptualization, Resources, Data curation, Supervision, Funding acquisition, Investigation, Visualization, Methodology, Writing—original draft, Project administration, Writing—review and editing.

2.9 Conclusion

This chapter describes our work in humanizing, plantizing and bacterializing evolutionarily related genes in yeast. The first major result is that despite much greater evolutionary divergence, the earlier finding that roughly half of genes can be replaced by cross-species orthologs[8] holds – despite bacteria being almost 4 times more distant to yeast than humans. In fact, after accounting for certain general peculiarities of bacterial genetics, more than half of the genes tested could be bacterialized. Among these, the heme biosynthesis pathway has proven strikingly universal: It can be effectively complemented by genes from all three tested species. Moreover, we observed a very distinct color phenotype owing to protoporphyrin secretion from yeast cells as a result of disruptions to pathways flux. The phenotype is readily quantifiable with chemical means, and is also apparent to even the naked eye. The step which generated this phenotype is the final iron chelation reaction, which is of particular interest: Mutations in it are associated with porphyria in humans, therefore the pink humanized yeast is potentially an excellent model for studying mutations and drugs of human porphyria in the much more experimentally tractable yeast system. Moreover, because heme is a crucial precursor to production of chlorophyll, a similar system could be investigated for studying herbicides.

2.9.1 Diverse Biochemistry of Heme Production

The results from the replacement of the first step of heme biosynthesis are particularly remarkable. While most of the pathway involves identical intermediates across species, the first step is notably different in plants and bacteria as compared to yeast and humans[54]. In all organisms the first step produces delta-aminolevulinic acid (ALA) which then becomes the substrate of the second step (ScHEM2, EcHemB, At-HEMB1/2 and HsALAD). Two distinct pathways are known for the production of ALA: The first, often called the "Shemin pathway", involves the condensation of succinyl-CoA and glycine, which is done by 5-aminolevulinate synthase HEM1 in yeast and ALAS1/2 in humans. In plants and most bacteria (including *E. coli*) the C5 pathway is used instead, which involves reduction of glutamyl-tRNA$^{\mathrm{Glu}}$ (Glu-tRNA) to glutamate-1-semialdehyde (GSA) by Glu-tRNA reductase (GluTR), which is encoded by EcHemA in *E. Coli* and AtHEMA1 in *A. thaliana*. GSA is converted to ALA by glutamate-1-semialdehyde-2,1-aminomutase (GSAM). Structurally, GSAM is very similar to ALAS – indeed, it is thought that ALAS evolved from GSAM. Thus GluTR is the more unique enzyme, responsible for the use of Glu-tRNA as the ultimate starting point for heme production.

Because the C5 pathway is not active in yeast, it is surprising that HEM1 could be bacterialized or plantized as it did. In bacteria, there are separate Glu-tRNAs used for protein elongation and heme biosynthesis. The latter bears $A_{53} \cdot U_{61}$ basepair where the former bears a conserved $G \cdot C$ pair[97] (notably *E. coli* is an exception). GluTR can recognize such features with high

specificity. Therefore, it is on the one hand surprising that presumably unmodified yeast Glu-tRNA was compatible with bacterial HemA, but somewhat explained in that the *E. coli* GluTR is slightly less specific.

We can suspect that introducing C5 activity to yeast would might have deleterious effects in that it creates an unnatural shunt of scarce Glu-tRNA. Indeed, Glu-tRNA starvation can have significant impact on the fitness of yeast, such as reducing the protein protein synthesis by 10-fold[98]. However, our bacterialized yeasts are not actually the first case of C5 activity in yeast: HemA has been expressed before (albeit without deleting HEM1). As such, the fact that yeast can meet the additional demand for Glu-tRNA imposed by GluTR activity is, though surprising, corroborated by literature. Moreover, HemA yeast is also not the first example of both pathways being active in a non-plant eukaryote: *Euglena* exhibits this phenomenon naturally[99]. Interestingly, *Euglena* heme production is separated by final purpose: The heme destined to mitochondria derives from the Shemin pathway while the heme destined to plastids derives from the C5-pathway.

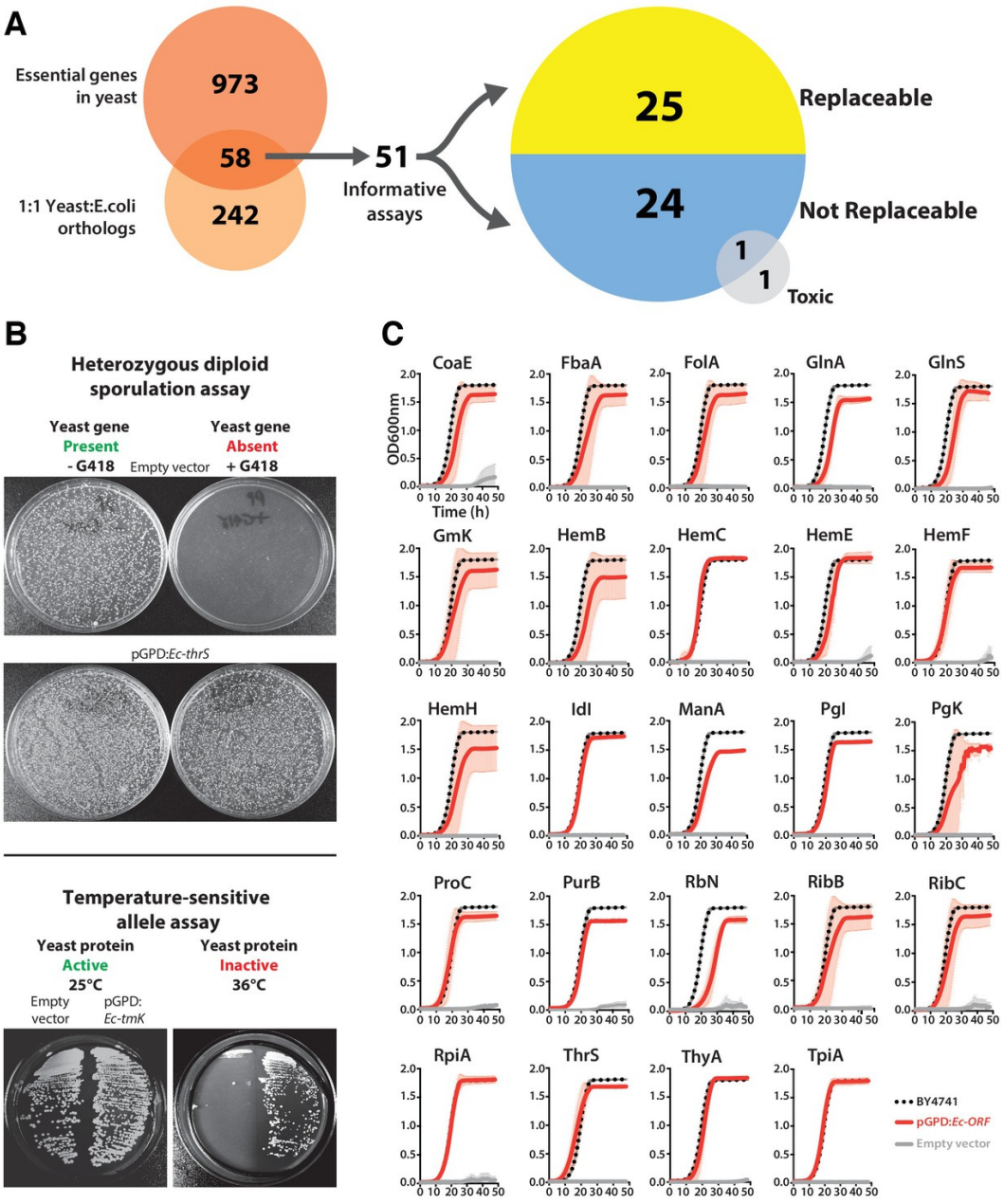

Figure 2.1: Systematic functional replacement of essential yeast genes by their human counterparts (Continued on next page.)

Figure 2.1: **(A)** Yeast and *E. coli* share hundreds of genes, 58 of which are essential in yeast and have clear 1:1 orthologs in either species. *E. coli* genes were cloned into a yeast expression vector under the control of a GPD promoter. 51 of these 58 *E. coli* genes provided informative assays for replaceability in yeast. Initial results from these complementation assays revealed that 25 of 51 (∼49%) *E. coli* genes could functionally replace their orthologous yeast counterparts. **(B)** Complementation assays were performed in two different yeast strain backgrounds, as shown for representative assays. In the case of a yeast strain with a temperature-sensitive allele of the yeast gene Sc-cdc8, cells carrying the empty vector control grow at the permissive-temperature (25 °C, yeast protein active) but not the restrictive-temperature (36 °C, yeast protein inactive), unlike cells expressing the *E. coli* ortholog (Ec-tmK), indicating that the *E. coli* gene can functionally replace the yeast gene. In the case of yeast heterozygous diploid (Sc-ths1Δ/Sc-THS1) deletion strain, cells are sporulated and haploid progeny grown on selective medium (-Ura-Arg-His-Leu+Can) in the absence (yeast gene present) or presence of G418 (200 µg/ml) (yeast gene absent). Cells expressing the *E. coli* ortholog (Ec-thrS) grow on G418-containing medium, unlike cells carrying the empty vector control, indicating successful complementation. **(C)** Haploid yeast gene deletion strains carrying plasmids expressing functionally replacing *E. coli* genes (red solid-lines) generally exhibit comparable growth rates to the wild type parental yeast strain BY4741 (black dotted-lines). The empty vector control (grey solid-line) showed no such growth rescue in the presence of G418. Mean and standard deviation plotted with N = 3.

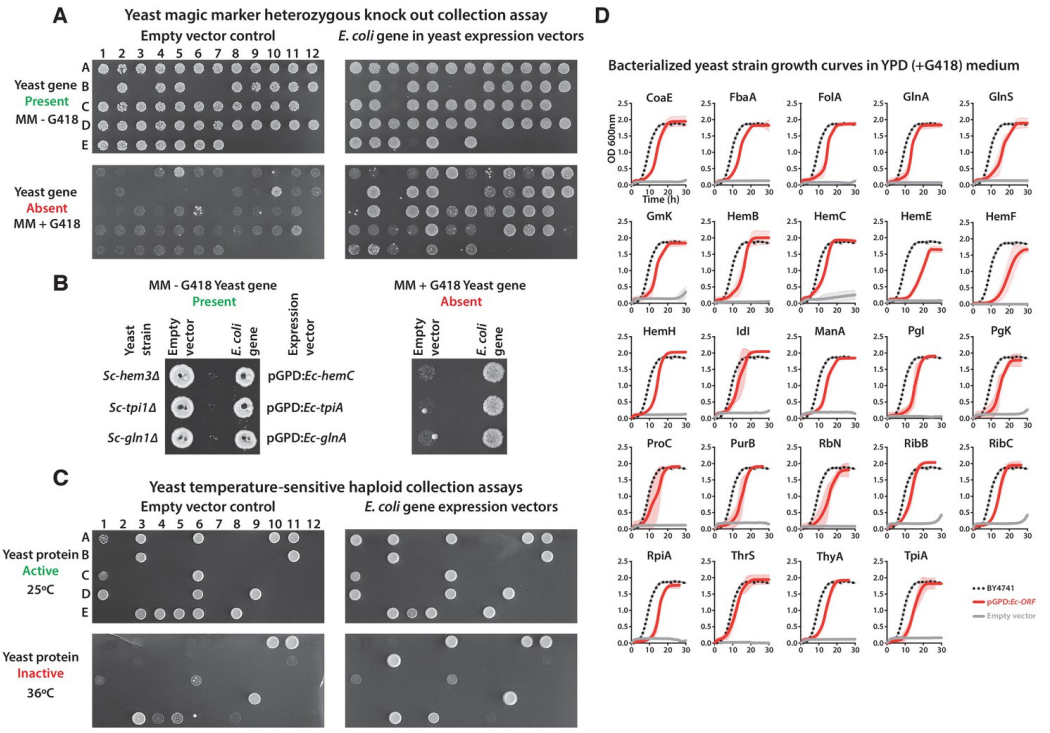

Figure 2.2: Complementation assays performed in a 96-well format in two different yeast strain backgrounds (Supplementary file 1) (A and **B)** Magic marker heterozygous diploid deletion yeast strains expressing *E. coli* genes were sporulated and the sporulation mix was spotted on magic marker agar medium (-Ura -Arg -His -Leu + Can) with (yeast gene absent) or without (yeast gene present) G418 (200 µg/ml). **(C)** Temperature-sensitive haploid yeast strains expressing *E. coli* genes grown at permissive temperature (25 °C) (yeast protein active) and at restrictive temperature (36 °C) (yeast protein inactive) on -Ura agar medium with G418 (200 µg/ml). Empty vector containing yeast cells were used as negative control for the experiment. **(D)** Haploid yeast gene deletion strains carrying plasmids expressing functionally replacing *E. coli* genes (red solid-lines) generally exhibit comparable growth rates to the wild type parental yeast strain BY4741 (black dotted-lines) as grown in YPD liquid medium in the presence of G418 (300 µg/ml). Mean and standard deviation plotted with N = 3.

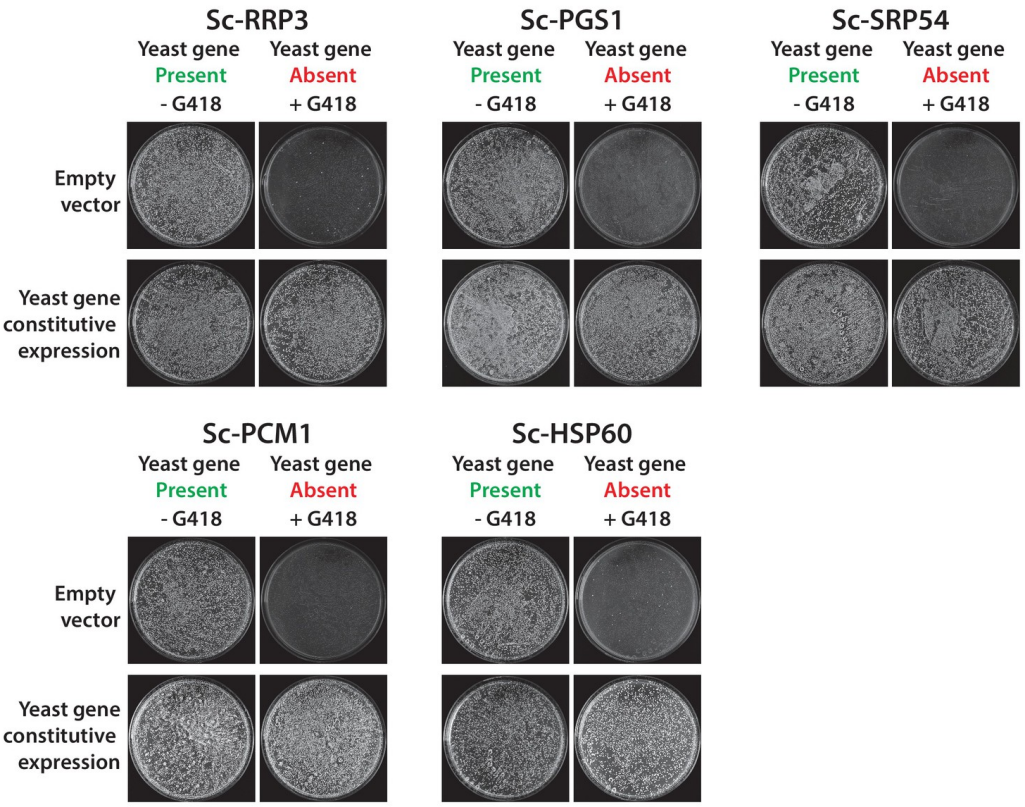

Figure 2.3: Constitutive plasmid expression of yeast genes efficiently replaced the corresponding genomic copies for 6 non-replaceable alleles. Bacterial orthologs of the yeast genes, Sc-RRP3, Sc-PGS1, Sc-SRP54, Sc-PCM1 and Sc-HSP60 did not show functional replacement when expressed from a constitutive GPD promoter. We expressed the corresponding yeast genes in a similar fashion under the control of the constitutive GPD promoter. All the tested yeast genes functionally replaced the corresponding yeast gene deletions. Empty vector containing yeast cells were used as negative control for the experiment.

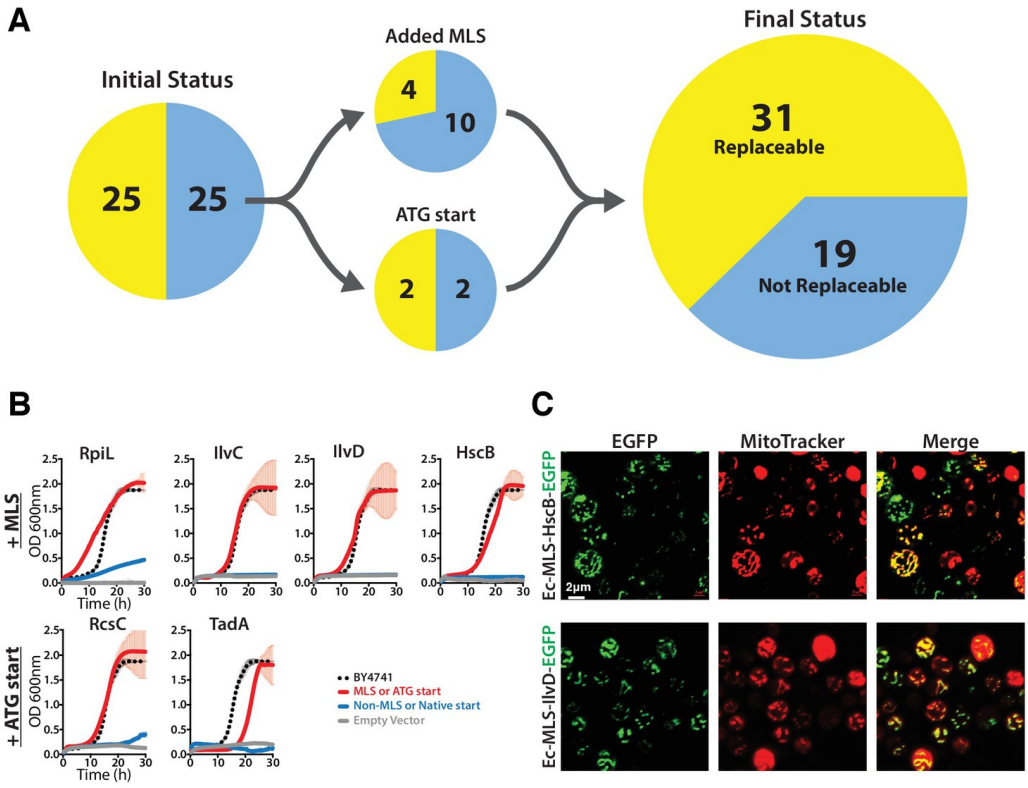

Figure 2.4: The addition of a mitochondrial localization signal (MLS) and mutation of start codons from GTG to ATG allows some *E. coli* genes to swap for their respective yeast orthologs.

Figure 2.4: (A) 14 of the 25 non-replaceable *E. coli* genes were predicted to function in mitochondria in yeast. 4 of 14 were replaceable after adding the MLS at the N-termini of the *E. coli* genes. Site-specific mutagenesis of *E. coli* gene start codon from GTG to ATG allowed two to functionally complement the corresponding yeast genes bringing the total number *E. coli* genes that functionally replace yeast genes to 31 of 51 ($\sim$61%). (B) Haploid yeast gene deletion strains carrying mitochondrially localized *E. coli* genes rescued the growth defect of the yeast gene (red solid-line) comparable to the wild type yeast (black dashed-line). The empty vector control (grey solid-line) and the yeast cells expressing of *E. coli* gene without MLS (blue-solid line) showed no such growth rescue in the presence of G418. Mean and standard deviation plotted with N = 3. (C) EGFP-tagged *E. coli* genes that functionally replaced the yeast gene function were imaged after MitoTracker red staining. EGFP-tagged Ec-MLS-HscB and Ec-MLS-IlvD (green) show colocalization with MitoTracker red stained mitochondria (red).

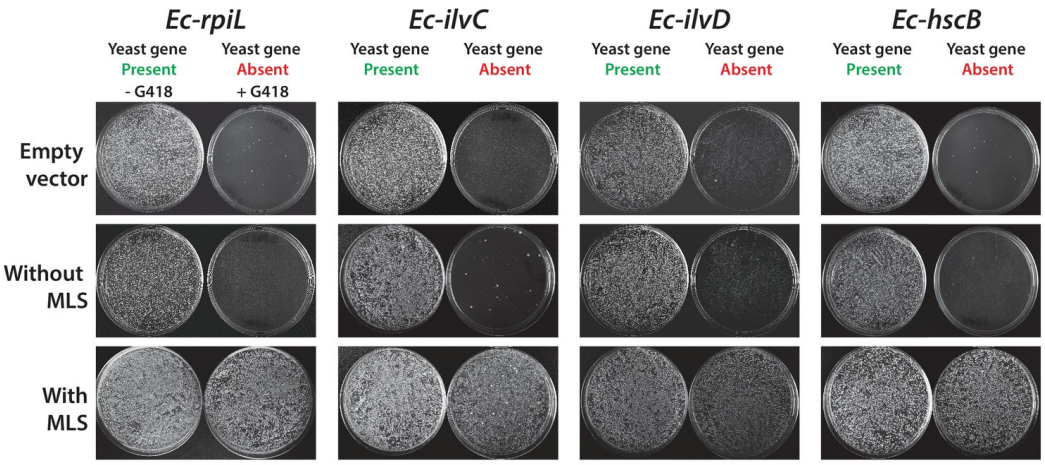

Figure 2.5: Some *E. coli* genes require a yeast mitochondrial localization signal to efficiently replace. The magic marker heterozygous diploid deletion yeast strains carrying empty vector or *E. coli* gene with or without MLS were sporulated and the sporulation mix was plated on magic marker agar medium (-Ura-Arg-His-Leu+Can) with or without G418 (200 µg/ml). *E. coli* genes Ec-rpiL, Ec-ilvC, Ec-ilvD and Ec-hscB without an appropriate mitochondrial localization signal cannot complement the corresponding yeast gene deletions Sc-mnp1, Sc-ilv5, Sc-ilv3 and Sc-jac1. However, expression of *E. coli* genes with yeast MLS efficiently rescued the growth defect of the corresponding yeast gene deletions.

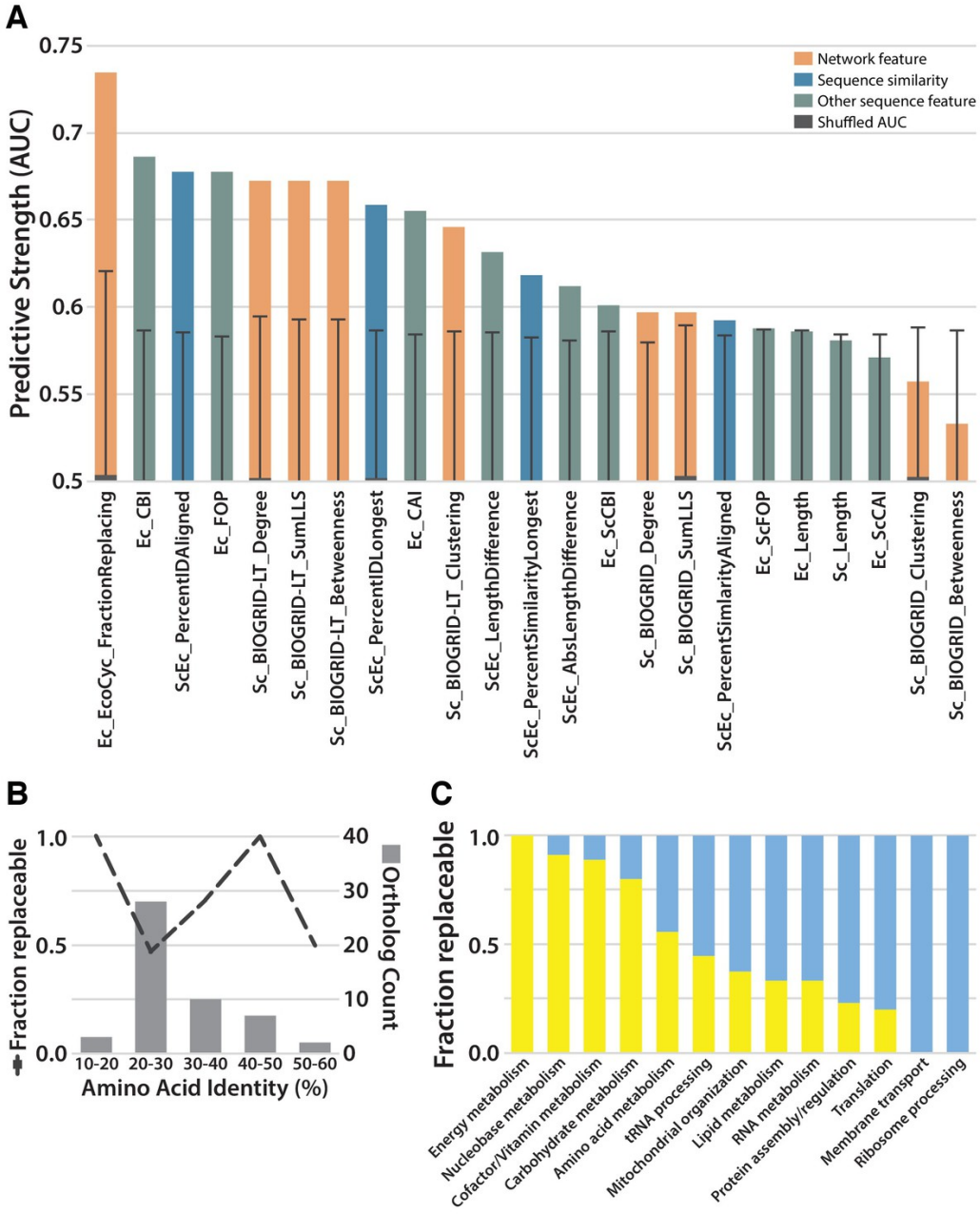

Figure 2.6: Replaceability of *E. coli* genes is a modular phenomenon. (Continued on next page.)

Figure 2.6: **(A)** Several quantitative properties of the tested genes were assessed for their ability to predict replaceability, measured as the area under a receiver operating characteristic curve (AUC). Having a high fraction of interaction partners that replace was the most predictive property tested, suggesting that the ability to replace is a modular phenomenon whereby genes functioning together are similarly able to replace. A Random Forest classifier constructed with all attributes boosted the maximum AUC to 0.79. **(B)** As shown in **(A)**, sequence similarity was not the most predictive feature. The fraction of replaceable genes in given ranges of similarity was variable, with the vast majority of orthologs being 20-30% identical, a range in which roughly half of proteins replaced. **(C)** Mapping of replaceability status onto yeast GO slim annotations revealed that GO categories have varying rates of replaceability, with core metabolic processes (e.g. energy metabolism, nucleobase metabolism) being largely replaceable while more specialized processes (e.g. protein assembly, membrane transport) were less so.

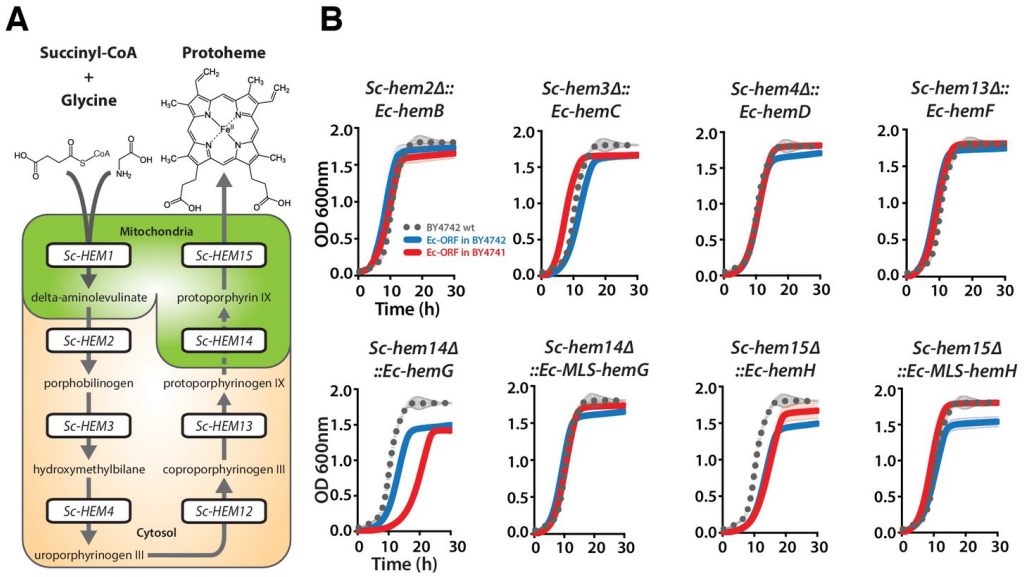

Figure 2.7: Bacterialization of yeast heme biosynthesis pathway genes at their native loci. **(A)** A schematic of the yeast heme biosynthesis pathway shows the beginning of the pathway in mitochondria using succinyl-CoA and glycine as precursors. The subsequent enzymatic reactions are cytosolic up until the penultimate and ultimate reactions which are mitochondrial. **(B)** Growth kinetics of CRISPR-Cas9 engineered yeast heme biosynthesis pathway genes replaced with the corresponding bacterial genes at their native yeast loci show efficient replaceability in both BY4741 (red solid-line) and BY4742 (blue solid-line) yeast strains. The wild type BY4741 growth curve is shown as a comparison (black dotted-line). Mean and standard deviation plotted with N = 3.

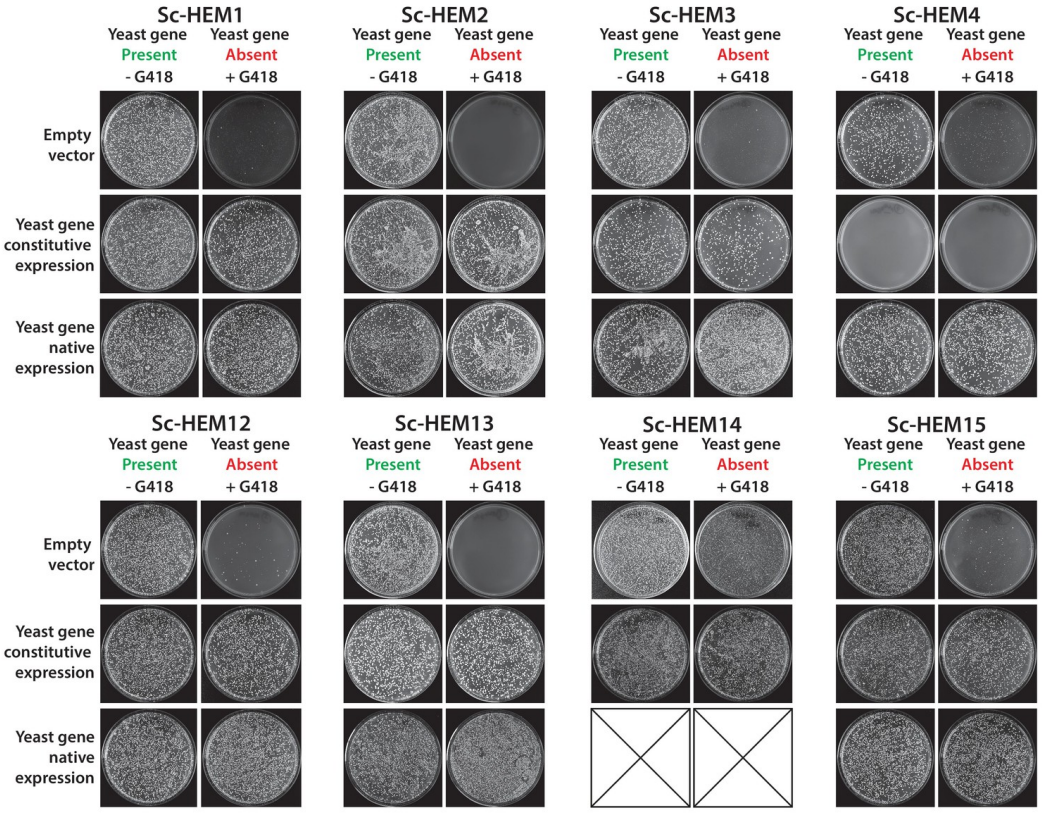

Figure 2.8: Constitutive or native plasmid-based expression of the yeast heme biosynthesis genes generally efficiently complemented growth defects in the corresponding yeast gene deletion strains. Heterologous expression of yeast genes Sc-HEM1, Sc-HEM2, Sc-HEM3, Sc-HEM4, Sc-HEM12, Sc-HEM13, Sc-HEM14 and Sc-HEM15 under the control of constitutive GPD promoter or native promoter efficiently rescued the growth defect of the corresponding yeast gene deletions respectively except in the case of Sc-HEM4. Sc-HEM4, when expressed constitutively, resulted in toxicity in the presence of the yeast gene at the native locus and did not complement the function in the absence of the yeast gene. This toxicity was relieved when the yeast gene was expressed under the control of the native yeast promoter.

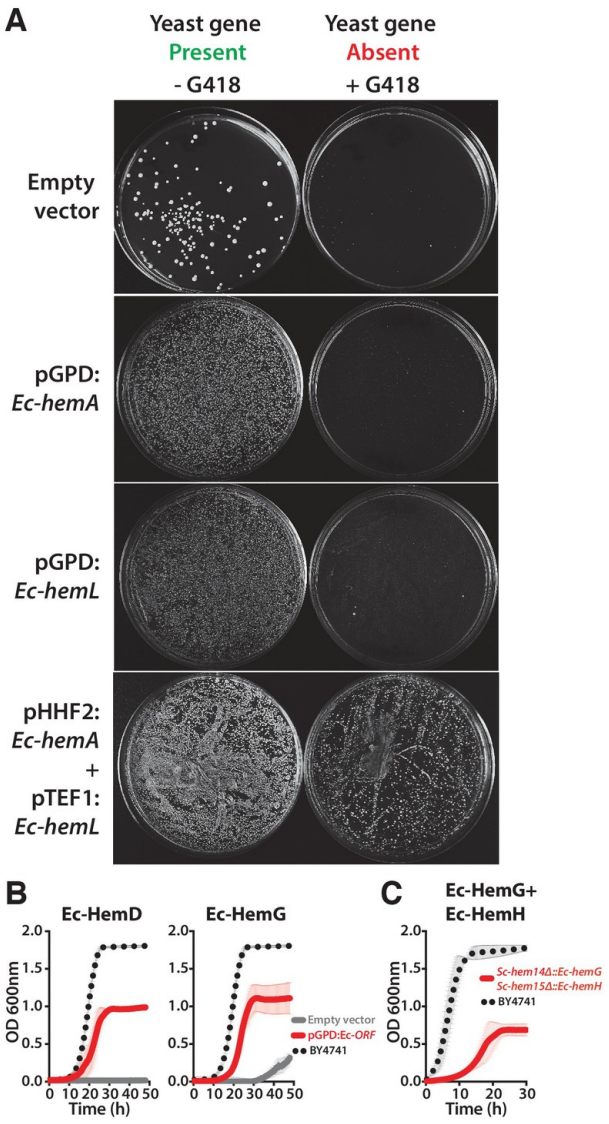

Figure 2.9: Ec-hemA and Ec-hemL carry out the initial reaction in *E. coli* heme biosynthesis and are both required to complement Sc-HEM1 deletion in yeast, and non-orthologous yeast genes are replaced by *E. coli* genes that carry out the identical reaction. (Continued on next page.)

Figure 2.9: **(A)** Expression of heme pathway genes of *E. coli*, Ec-hemA or Ec-hemL, individually cannot complement the lethal growth defect of the deletion of Sc-HEM1 gene in yeast. Co-expression of Ec-HemA and Ec-HemL efficiently rescued the growth defect of Sc-hem1 gene deletion in yeast. **(B)** Growth curves of yeast strains with deletions of Sc-hem4 and Sc-hem14 genes (grey solid-line) show functional replaceability (red solid-line) by the non-orthologous *E. coli* genes Ec-hemD and Ec-hemG that carry out identical enzymatic reactions to the corresponding yeast genes. The wild type BY4741 growth curve is shown as a comparison (black dotted-line). The empty vector control (grey solid-line) showed no such growth rescue in the presence of G418. **(C)** Growth curve of engineered yeast strain Sc-hem14Δ::Ec-hemG; Sc-hem15Δ::Ec-hemH in YPD medium harboring *E. coli* genes at the native yeast loci. The strain displayed a growth defect (red solid-line) compared to the wild type BY4741 strain (black dotted-line). Mean and standard deviation plotted with N = 3.

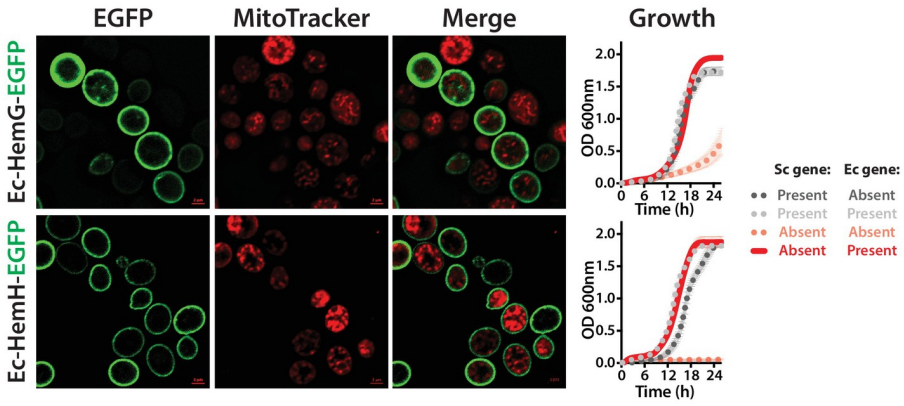

Figure 2.10: The penultimate and ultimate heme pathway enzymes in yeast are replaceable by their bacterial orthologs, in spite of mis-localizing to the plasma membrane. EGFP-tagged Ec-HemG and Ec-HemH localize to the plasma membrane in yeast. The EGFP-tagged proteins do not localize to the mitochondria since no clear co-localization is observed with the Mitotracker red stain. EGFP-tagged Ec-HemG and Ec-HemH expression (red solid-line) efficiently rescue the growth defects of the respective yeast gene deletions (Sc-hem14 and Sc-hem15) (pink dotted-line) comparable to the wild type yeast (black dotted-line). Empty vector control is incapable of rescuing the growth defect of the deletion strains (grey dotted-line).

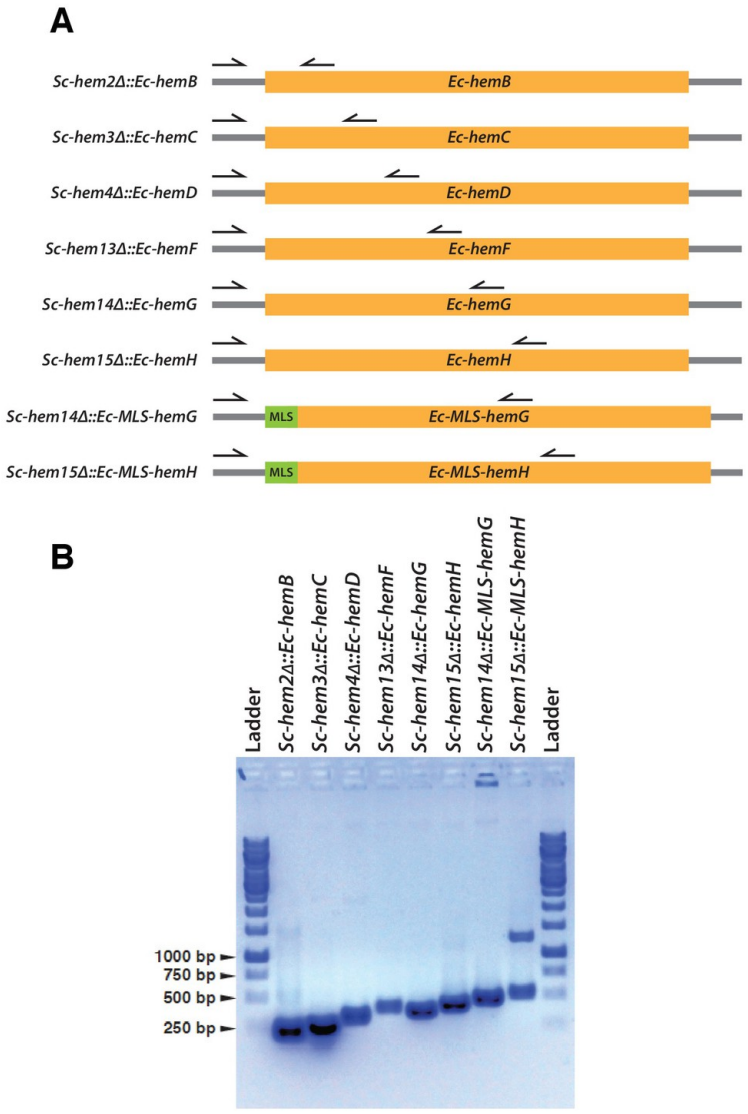

Figure 2.11: Confirmation of CRISPR-Cas9 mediated bacterialized yeast strains. **(A)** Schematics of the yeast heme biosynthesis pathway gene loci carrying functionally replaceable *E. coli* genes while retaining their native promoters and terminators. The arrows indicate the primers used to confirm the replacement (refer to Supplementary file 3). **(B)** PCR amplification of expected size was obtained for each individual bacterialized yeast strains.

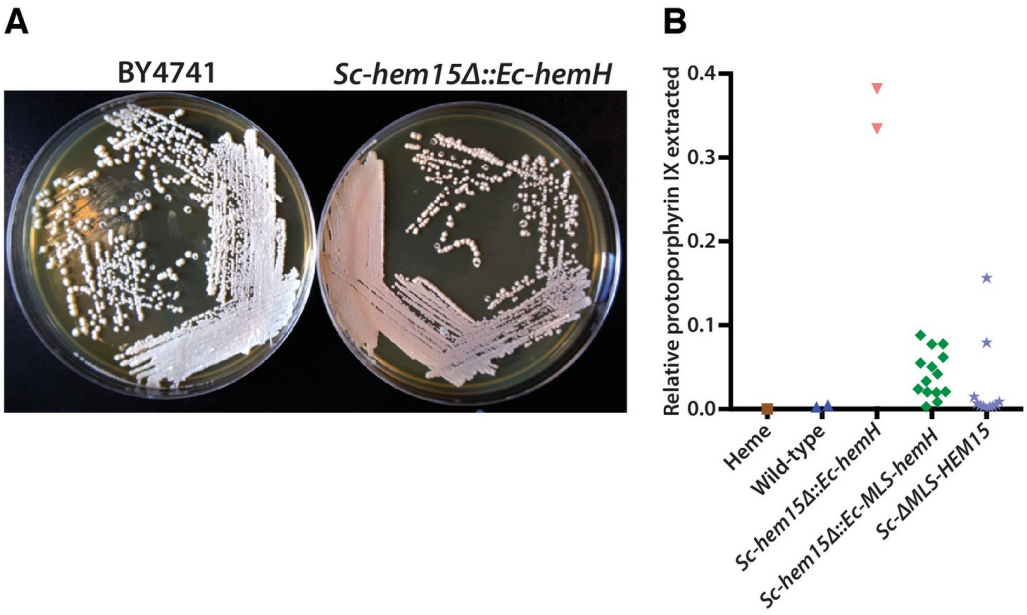

Figure 2.12: Mislocalization of the bacterialized ferrochelatase enzyme identifies a porphyria-like phenotype in yeast. **(A)** Bacterialization of the ultimate yeast gene in the heme biosynthesis pathway results in a distinct pink colony phenotype on YPD agar medium. In contrast, wild type BY4741 strain colonies appear as creamy-white. **(B)** Acetate-extracted secreted products from the pink Sc-hem15Δ::Ec-hemH strains show strongly enhanced fluorescence at 635 nm (excitation 399 nm), comparable to a protoporphyrin IX standard and unlike a heme standard or extracts from the parental BY4741 strain. The introduction of an MLS to the bacterialized yeast strain (Sc-hem15Δ::Ec-MLS-hemH) significantly reduced protoporphyrin IX secretion, while deletion of the MLS from the native yeast locus in strain Sc-ΔMLS-HEM15 caused several strains to increase protoporphyrin IX secretion.

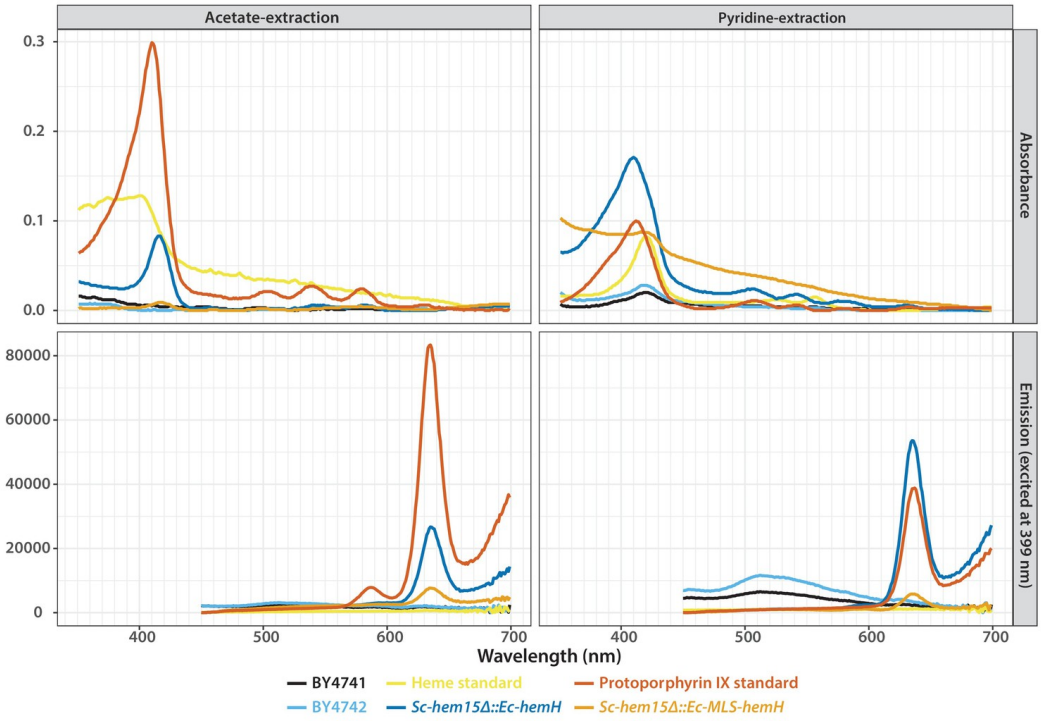

Figure 2.13: Absorbance (top) and emission (bottom) spectra of extracts obtained from acetate (left) and pyridine (right) extraction of the wild type or bacterialized yeast colonies grown on YPD medium. Purified protoporphyrin IX (red solid-line) or heme (yellow solid-line) were used as standards. Extract from the bacterialized Sc-hem15Δ::Ec-hemH yeast strain (dark blue-line) matched with that of the protoporphyrin IX standard. Bacterialized ScHEM15Δ::Ec-MLS-hemH yeast strain (orange solid-line) showed significantly reduced peak for protoporphyrin IX. Extracts from wild type BY4741 (black-line) and BY4742 (light blue solid-line) were used as controls.

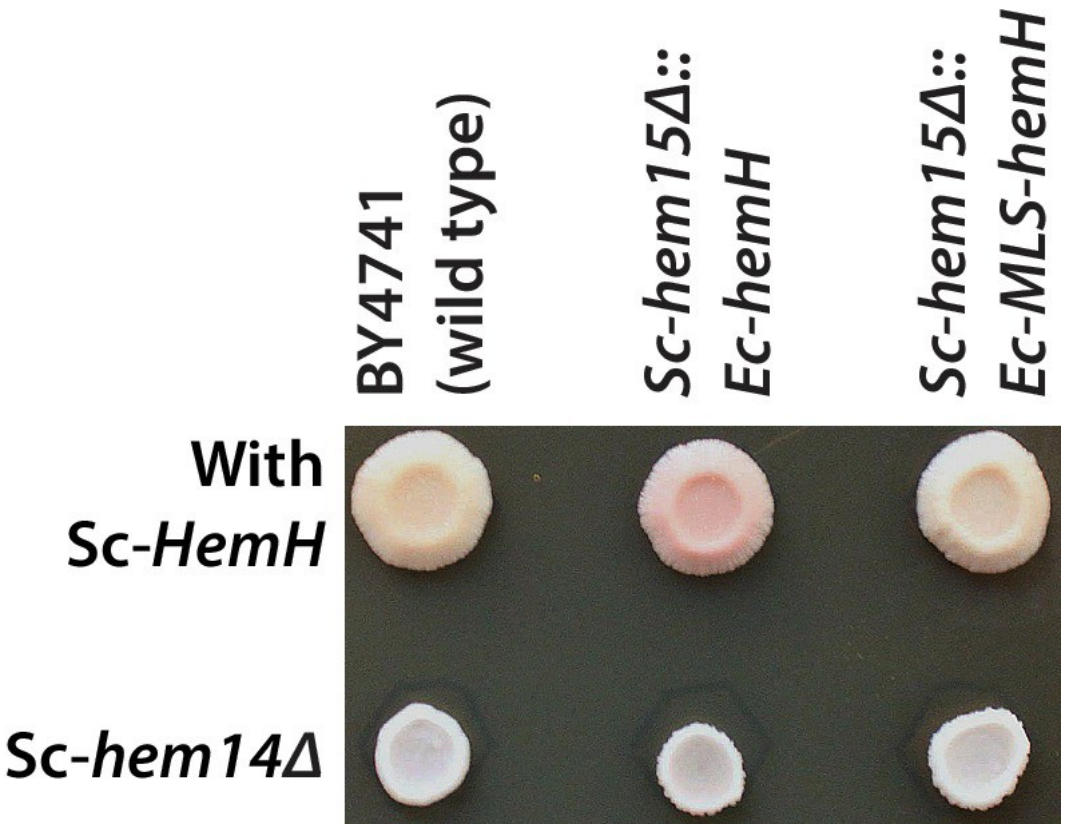

Figure 2.14: Deletion of protoporphyrinogen oxidase, Sc-HEM14, in the Sc-hem15Δ::Ec-hemH strain suppressed the porphyria-like pink phenotype. Top row from left show growth spots of the BY4741 wild type, Sc-hem15Δ::Ec-hemH and Sc-hem15Δ::Ec-MLS-hemH yeast strains. Bottom row from left show corresponding strains harboring Sc-hem14 deletion.

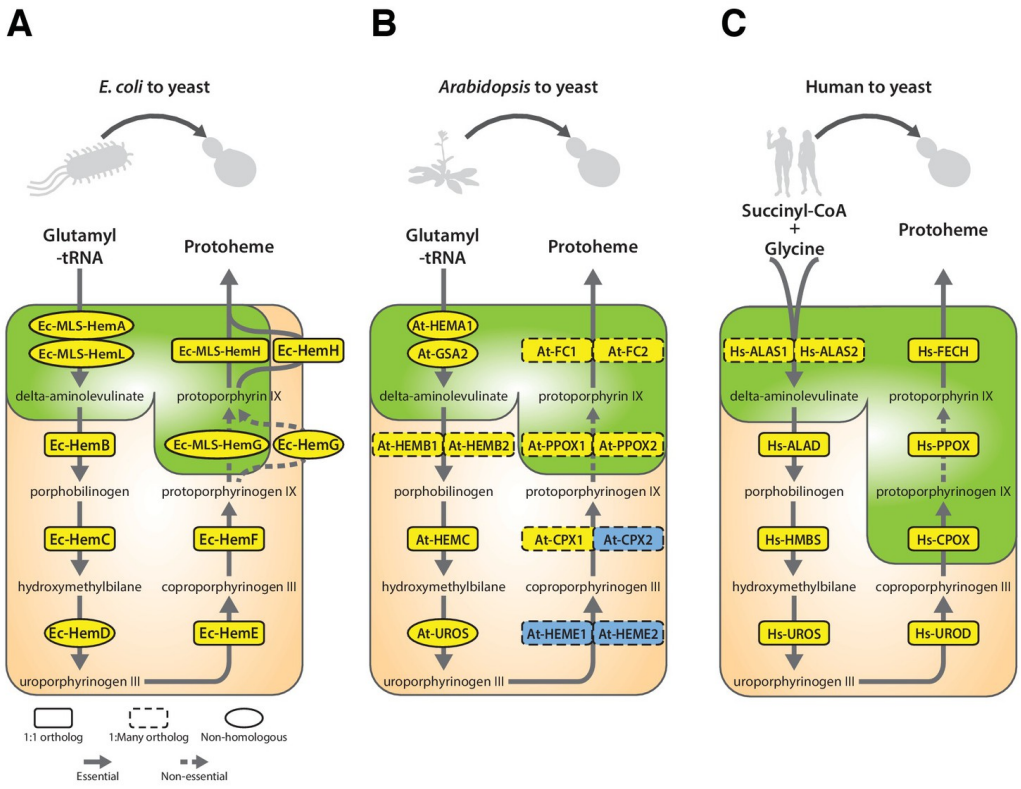

Figure 2.15: Yeast heme biosynthesis pathway enzymes can be successfully replaced by orthologs or analogs from bacteria, plants, and humans, in spite of alterations to subcellular localization. Enzymatic steps of extant bacterial and eukaryotic heme biosynthesis pathways are identical save for the starting metabolites and conversion to delta-aminolevulinate; bacteria also exhibit non-orthologous gene displacement of several enzymes. Heme biosynthesis occurs in the bacterial cytoplasm and inner membrane, the human and yeast in mitochondria and cytoplasm, and the plant in chloroplast and cytoplasm. In spite of these localization changes over evolution, most of the defects in growth rate and viability conferred by heme pathway mutations in yeast can be complemented by introduction of the corresponding **(A)** bacterial genes, **(B)** plant genes (except for At-HemE), and **(C)** human genes. Yellow indicates a replaceable gene, blue non-replaceable.

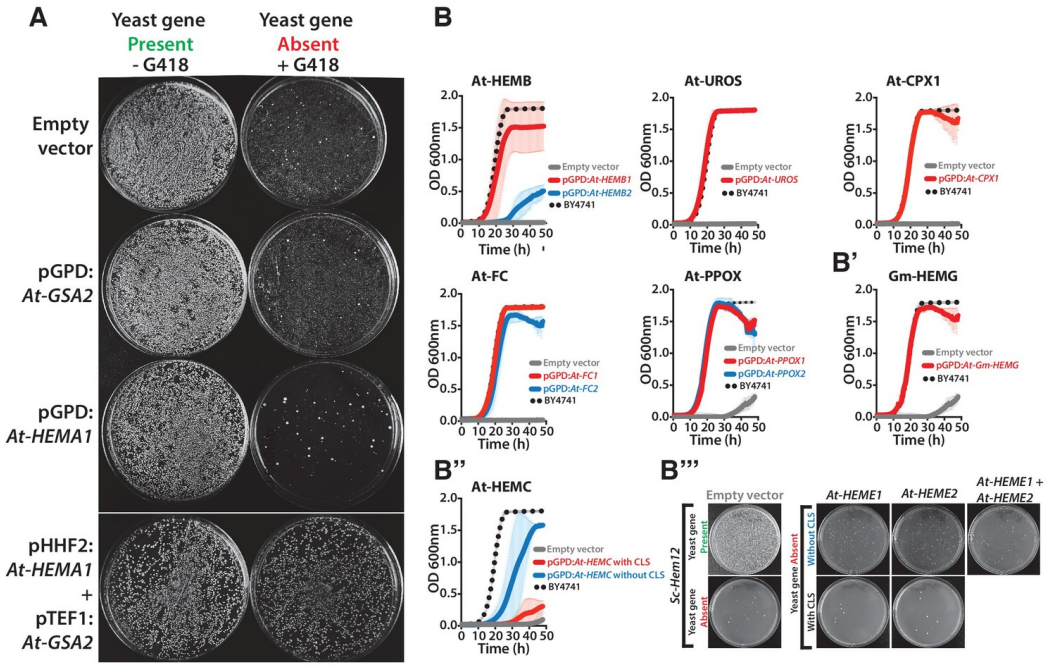

Figure 2.16: Heme biosynthesis genes from *Arabidopsis thaliana* and Glycine max generally efficiently replace their counterparts in yeast, except in the case of ΔSc-Hem12.

Figure 2.16: **(A)** Expression of heme pathway genes from *Arabidopsis thaliana*, At-HEMA1 or At-GSA2, individually cannot complement the lethal growth defect of the deletion of Sc-hem1 gene in yeast. Co-expression of At-HEMA1 and At-GSA2 rescued the growth defect of Sc-hem1 gene deletion in yeast. **(B)** Haploid yeast gene deletion strains carrying plasmids expressing functionally replacing *Arabidopsis* (red or blue solid-lines) and **(B')** Glycine max (Gm-HEMG) heme pathway genes (red solid-line) generally exhibit comparable growth rates to the wild type parental yeast strain BY4741 (black dotted-line) as grown in magic marker liquid medium in the presence of G418 (200 µg/ml). (B") Native At-HEMC with chloroplast localization signal (CLS) showed poor replaceability in yeast (red solid-line). Removal of the CLS from At-HEMC allowed efficient rescue of the corresponding yeast gene deletion, ΔSc-Hem3 (blue solid-line). (B"') However, neither the expression of Arabidopsis proteins At-HEME1 or At-HEME2 (with or without CLS) alone nor their co-expression could functionally rescue the corresponding yeast gene deletion, ΔSc-Hem12. Wild type BY4741 haploid strain is plotted for comparison (black dotted-line). Strains carrying empty vector were used as controls (grey solid-line). Mean and standard deviation plotted with N = 3.

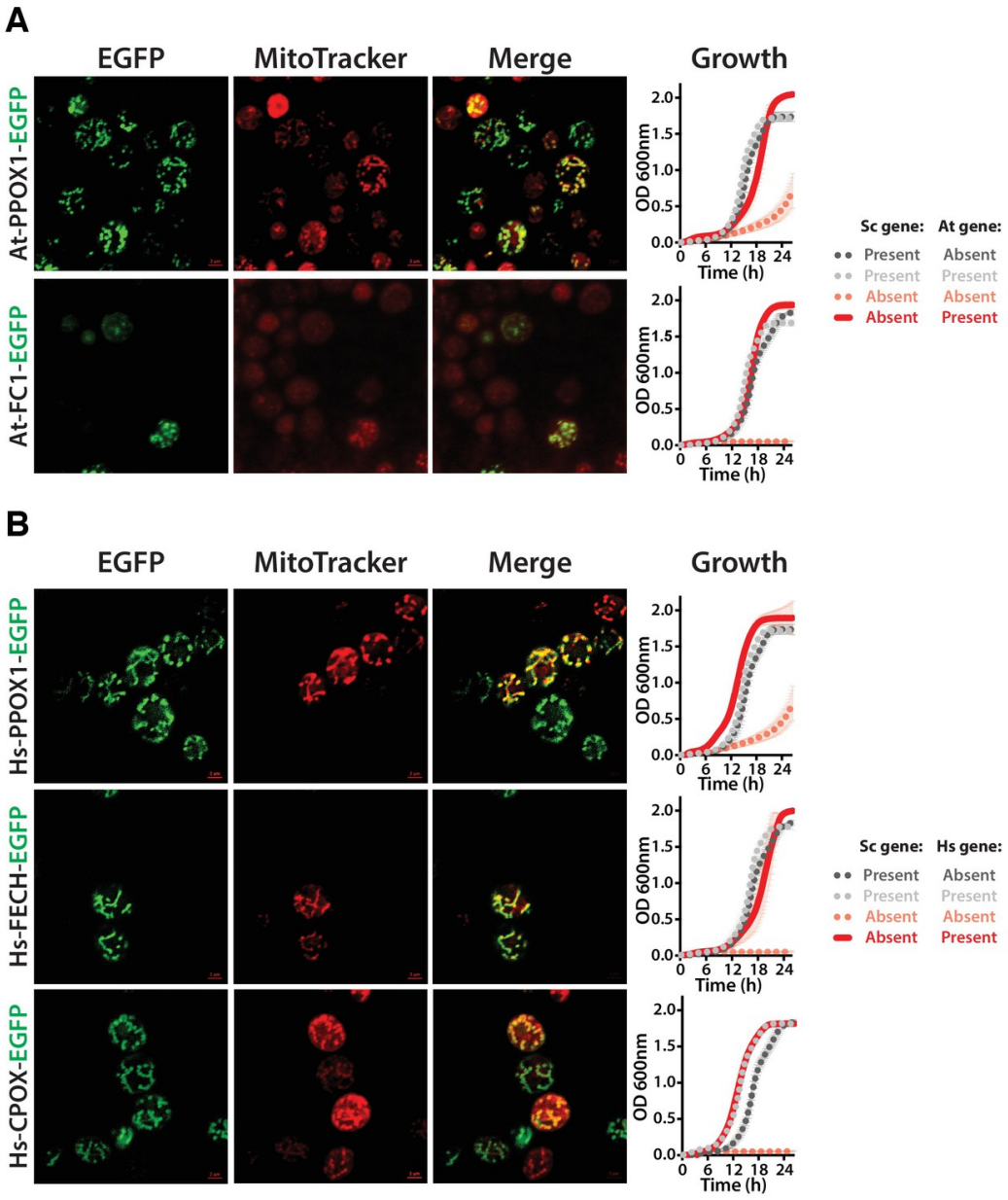

Figure 2.17: Heme biosynthesis enzymes normally localized to plant chloroplasts or human mitochondria localize to the mitochondria when expressed in yeast.

Figure 2.17: (A) EGFP-tagged penultimate At-PPOX1-EGFP and ultimate At-FC1-EGFP proteins localize to mitochondria in yeast. Green fluorescence proteins co-localized with Mitotracker red-stained mitochondria. In certain cases, At-FC1-EGFP formed aggregates. Expression of EGFP-tagged plant genes, At-PPOX1-EGFP and At-FC1-EGFP (red solid-line), efficiently rescue the growth defect of the corresponding yeast gene deletions (pink dotted-line). The over-expression of the tagged proteins is not toxic to the wild type yeast strain (grey dotted-line). The growth rescue by plant genes is as efficient as the wild type BY4741 yeast strain (black dotted-line). Mean and standard deviation plotted with N = 3. **(B)** The EGFP-tagged last three heme pathway genes from humans localize to mitochondria in yeast. The green fluorescence co-localized with the Mitotracker red-stained mitochondria in yeast. Expression of EGFP-tagged human genes, Hs-PPOX-EGFP, Hs-FECH-EGFP and Hs-CPOX-EGFP (red solid-line), efficiently rescue the growth defect of the corresponding yeast gene deletions (pink dotted-line). The over-expression of the tagged proteins is not toxic to the wild type yeast strain (grey dotted-line). The growth rescue by the human genes is as efficient as the wild type BY4741 yeast strain (black dotted-line). Mean and standard deviation plotted with N = 3.

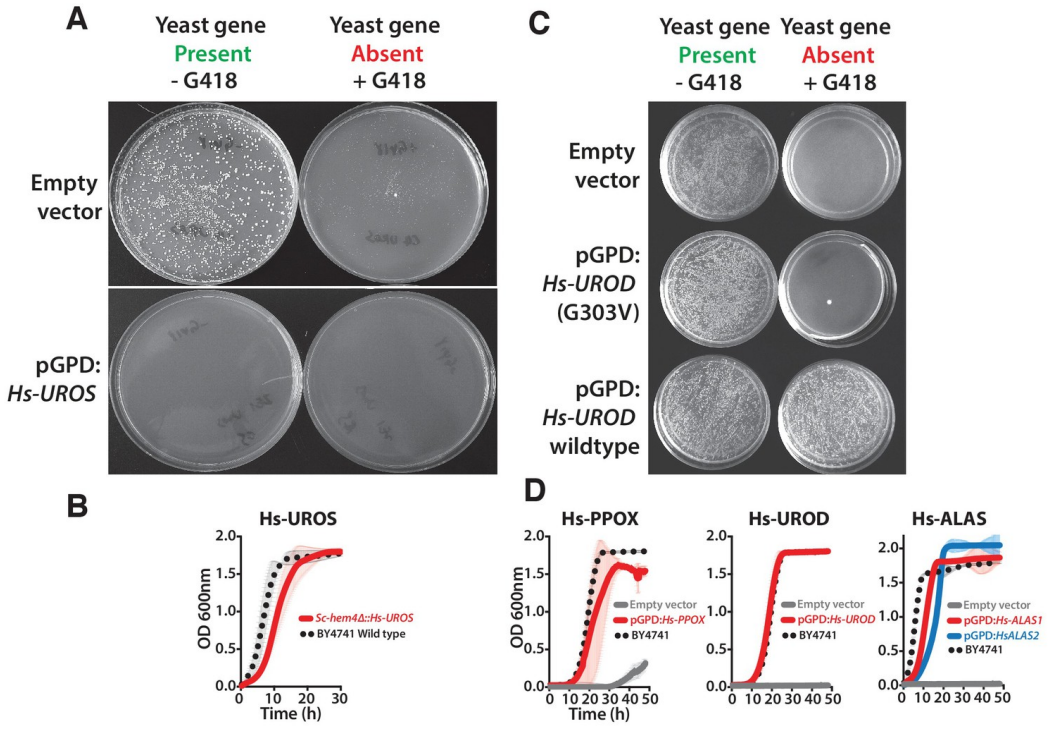

Figure 2.18: Human heme biosynthesis genes efficiently replace their yeast counterparts.

Figure 2.18: Functional replacement of human genes in yeast. **(A)** Expression of Hs-UROS in Sc-hem4 heterozygous diploid deletion yeast strain resulted in toxicity post-sporulation as seen by the lack of growth on either magic marker agar medium with (yeast gene present) or without G418 (yeast gene absent). **(B)** This toxicity was relieved by replacing the human Hs-UROS at the native yeast locus. Growth curve of the humanized yeast Sc-hem4::Hs-UROS strain (red-solid line) showed comparable growth to the wild type yeast BY4741 (black dotted-line). **(C)** Expression of human Hs-UROD (a human orfeome clone with G303V mutation) in Sc-hem12 heterozygous diploid deletion yeast strain did not complement the growth defect of the yeast gene as shown by plating the post sporulation mix on magic marker medium with or without G418. Reverting the sequence to the wild type Hs-UROD gene resulted in efficient rescue of the growth defect of the corresponding yeast gene. **(D)** Expression of human genes, Hs-PPOX, Hs-UROD, Hs-ALAS1 (red solid-line) and Hs-ALAS2 (blue solid-line), efficiently rescue the growth defect of the corresponding yeast gene deletions (grey solid-line), Sc-hem14 and Sc-hem1, respectively. The rescue was largely comparable to the wild type BY4741 yeast strain (black dotted-line). Strains carrying empty vector were used as controls (grey solid-line). Mean and standard deviation plotted with N = 3.

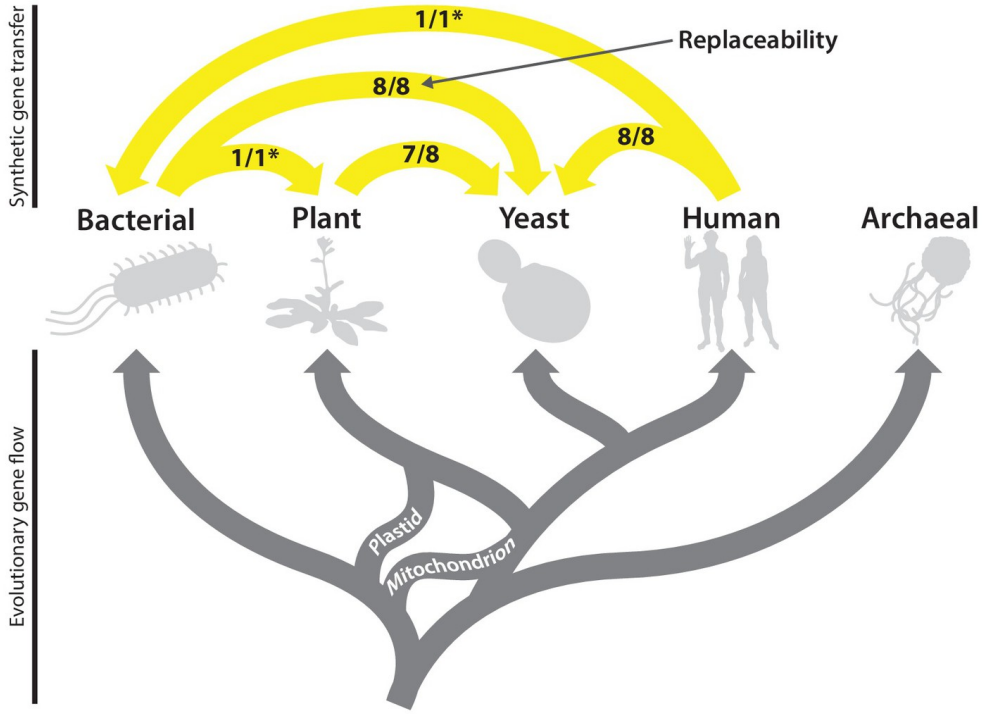

Figure 2.19: The complex evolutionary history of the heme biosynthesis pathway is reflected in high replaceability across species. In eukaryotes, heme biosynthesis enzymes have been replaced historically by endosymbiosis events from bacteria, leading to higher similarity across these lineages, while the archaeal pathway appears to be more divergent[100]. Following the endosymbiosis of the cyanobacterial chloroplast, plants adopted most of the chloroplast-derived heme biosynthesis genes, losing many ancestral eukaryotic heme pathway genes[73]. Yeast and humans both retain the predicted ancestral eukaryotic heme biosynthesis pathway. While enzymatic steps are mostly shared between yeast, plants, bacteria, and humans, localization of individual proteins differs substantially between species. Asterisks indicate results curated from literature.

Chapter 3

Single-step Precision Genome Editing in Yeast Using CRISPR-Cas9

In my initial foray into yeast humanization, a major point of interest was the demand for sophisticated methods for probing replaceability. A very effective method was to drive the expression of the complementing human ortholog from a plasmid, in a deletion strain. This approach had the advantage of being straightforward, feasible at the time, and being able to leverage existing deletion collections. The promoters in the plasmid were constitutive. Consequently, a potential confound is introduced into every experiment, because native transcriptional regulation is impaired and the gene is overexpressed.

A second method was knock-in/knock-out strategy based on first inserting a selectable marker cassette into the target locus, and then using elements embedded in the cassette to induce HR and implement the desired sequence at that locus. For essential genes, this necessitated the use of temporary complementing plasmids also. The state of the art in this area was developed by Dr. Chris Yellman in our lab in the form of the IT-series of cassettes, which I

This chapter was previously published in Akhmetov A, *et al.* (2018) *Bio Protoc.*, 8(6):e2765. I contributed to the development of the technique, experiments, analysis and the writing of the manuscript.

have also used in my initial experiments. However, the cassette-based strategy is somewhat time and labor intensive.

One of my first tasks as a graduate student was to identify a better technology to use for replacements. After some research, CRISPR emerged as a promising candidate. At the time, there were very few reports of CRISPR activity in yeast, and I made considerable progress in optimizing the elements know into a practical CRISPR-based gene replacement system. As other groups working in parallel on similar themes published their findings, I was able to synthesize their discoveries with mine and produce a very versatile, single-step, universal gene editing strategy based on CRISPR technology. This enabled much of the work in our eLife publication[37], and afterward, we were invited to write a detailed protocol which is reproduced below (with minor formatting changes).

This work was eventually published as [21]. Soon after publication, we were serendipitously informed that another group had succeeded in reproducing our results. Notably, though the protocol is written with a focus on humanization, it is in principle equally well suited to any arbitrary genetic editing. Due to the inherent DNA cleavage activity of CRISPR, the strains produced by this protocol can be used as gene-drives with super-Mendelian inheritance of the resulting edit in subsequent generations.

3.1 Abstract

Genome modification in budding yeast has been extremely successful largely due to its highly efficient homology-directed DNA repair machinery. Several methods for modifying the yeast genome have previously been described, many of them involving at least two-steps: insertion of a selectable marker and substitution of that marker for the intended modification. Here, we describe a CRISPR-Cas9 mediated genome editing protocol for modifying any yeast gene of interest (either essential or nonessential) in a single-step transformation without any selectable marker. In this system, the Cas9 nuclease creates a double-stranded break at the locus of choice, which is typically lethal in yeast cells regardless of the essentiality of the targeted locus due to inefficient non-homologous end-joining repair. This lethality results in efficient repair via homologous recombination using a repair template derived from PCR. In cases involving essential genes, the necessity of editing the genomic lesion with a functional allele serves as an additional layer of selection. As a motivating example, we describe the use of this strategy in the replacement of HEM2, an essential yeast gene, with its corresponding human ortholog ALAD.

3.2 Background

Saccharomyces cerevisiae (Bakers yeast) has a long history as a genetically tractable organism, and there are an array of methodologies to manipulate the yeast genome. However, until recently it has been necessary to apply selection to isolate clones possessing the desired genetic alteration [96, 20, 23, 37]. In cases where arbitrary, scar-less editing of the genome is desired, the solution is typically a two-step process: First a selectable cassette (containing the URA3 marker, for example), flanked by homology arms targeting the region of interest, and sometimes containing nuclease targeting sites (i.e., I-SceI sites) to aid in the removal of the cassette at the later stage, is knocked in via homologous recombination (HR). The small subpopulation of successful integrants is isolated by selecting for the cassette. Second, the marker is eliminated through highly efficient sequence specific methods such as site-specific recombination or endonuclease cleavage (I-SceI) to generate the desired form of the edited genomic locus. Two steps are necessary because no method was available which is both scar-less and efficient enough such that no selection is required.

The development of CRISPR/Cas9 technology in yeast has eliminated the need for this two-step process. Cas9 efficiently creates double-stranded breaks (DSBs) in yeast DNA at virtually any arbitrary locusprovided a PAM sequence is proximal to the desired cut site. When an appropriate repair template is provided, these DSBs are repaired through the endogenous HR system of yeast. Cas9 directed to the desired genomic locus via the guide RNA se-

quence creates double-stranded break (DSB) in the genome. The CRISPR target site is retained in cells which fail to repair the target site as expected, which allows Cas9 to repeatedly cleave the same region until HR-mediated editing takes place. Rarely, non-homologous end-joining (NHEJ) can generate mutations which block Cas9 cleavage despite failing to incorporate the expected genomic alterations. More commonly, cells simply succumb to the stress of repeated Cas9-induced genomic cleavages. In an appropriately conducted experiment, the majority of the surviving population tends to be cells which have lost their CRISPR target site by incorporating the desired genomic alteration via HR. Cas9 thus acts as a counter-selection acting directly on genomic sequence, rather than its phenotypic manifestations.

Here, we use an approach developed by Dueber and colleagues [23] to rapidly generate single, self-contained plasmids that express both the Cas9 nuclease and guide RNA required for targeting a desired locus. These plasmids, when co-transformed with an appropriate repair template provided as a linear PCR product, allow efficient, precise, single-step replacement of any arbitrary yeast gene with an introduced sequence of interest. Only selection for the Cas9 and gRNA-expressing plasmid is required, which tends to select for correct genomic modification by proxy due to efficiency of targeting and repair. This strategy was used extensively in our ortholog complementation research [37] to rapidly humanize, bacterialize and plantize many essential yeast genes. A CRISPR based approach is uniquely suited to this case, because it strongly encourages HR with functional alleles. False positives, arising from CRISPR sites

being mutated by NHEJ without incorporation of a new allele, are minimal because they are often not viable. Additionally, disruption of the target genes function is brief, eliminating the need for constructing and maintaining a complementing plasmid to sustain yeast through an otherwise lengthy engineering process. Further, given that CRISPR selects against sequence regardless of function, it is still possible and practical to alter non-essential genes (or even non-genic regions) with this technique; indeed, we have reported successful humanization of the non-essential yeast gene HEM14 with this method [37] and we have used this system to incorporate site-directed changes in proteins with high efficiency.

3.3 Materials and Reagents

1. Pipette tips (Mettler Toledo, catalog numbers: 17005872, 17005874, 17007089)
2. 96-well plate (VWR, catalog number: 82006-636)
3. 0.2 µmfilter (Fisher Scientific, catalog number: 09-719C)
4. Petri plates (VWR, catalog number: 25384-342)
5. Yeast (BY4741)
6. MoClo Yeast Toolkit (YTK, Addgene kit, Addgene, catalog number: 1000000061). Toolkit includes plasmids pYTK050, pYTK003, pYTK072, pYTK083, pYTK036, pYTK008, pYTK047, pYTK073, pYTK074, pYTK081 and pYTK084
7. PCR template for the sequence which will replace the target gene (e.g., cDNA, plasmid-based clone, etc.)
8. Note: For demonstration purposes, this protocol will assume replacement of *S. cerevisiae* HEM2 with its human ortholog ALAD.
9. NEB 5-alpha Competent *E. coli* (New England Biolabs, catalog number: C2987)
10. DNA stain (Thermo Fisher Scientific, Invitrogen^{TM}, catalog number: S33102)
11. T7 ligase (New England Biolabs, catalog number: M0318S)
12. T4 ligase buffer (New England Biolabs, catalog number: B0202S)
13. Restriction enzymes BsaI (New England Biolabs, catalog number: R0535S) and BsmBI (New England Biolabs, catalog number: R0580S)

14. LB plates with antibiotic selection
    (a) Ampicillin (Sigma-Aldrich, Roche Diagnostics, catalog number: 10835242001)
    (b) Spectinomycin (Sigma-Aldrich, catalog number: PHR1426)
15. Chloramphenicol (Sigma-Aldrich, catalog number: C0378)
16. High-fidelity DNA polymerase for repair template PCR, such as KAPA HiFi (Kapa Biosystems, catalog number: KK2601)
17. Zymo DNA Clean&Concentrator-25 kit (Zymo Research, catalog number: D4005)
18. Zymo EZ yeast transformation II kit (Zymo Research, catalog number: T2001)
19. Optional: 100 mM lithium acetate can be used in place of EZ 1 solution from the EZ competent yeast cell kit. (Lithium acetate can be obtained from Sigma-Aldrich, catalog number: L6883)
20. Accuprime Pfx (Thermo Fisher Scientific, Invitrogen^{TM}, catalog number: 12344024)
21. Optional: 5-fluoroorotic acid (Sigma-Aldrich, catalog number: F5013), if counter-selection will be used (see Procedure E)
22. D-Sorbitol (Sigma-Aldrich, catalog number: S3889)
23. Zymolyase (MP Biomedicals, catalog number: 320921)
24. LB Broth, Lennox (BD, catalog number: 240210)
25. YPD powder (BD, catalog number: 242820)
26. Agarose (Thermo Fisher Scientific, Invitrogen^{TM}, catalog number:

16500500)

27. Agar (SERVA Electrophoresis, catalog number: 11396)
28. Yeast nitrogen base without amino acids (BD, catalog number: 291940)
29. Ammonium sulfate (Sigma-Aldrich, catalog number: A4418)
30. Dextrose (Avantor Performance Materials, catalog number: 1919)
31. SC-Ura dropout powder (Sigma-Aldrich, catalog number: Y1501)
32. Zymolyase solution (see Recipes)
33. Lithium acetate (see Recipes)
34. LB medium (see Recipes)
35. YPD agar plates (see Recipes)
36. SD-Ura agar plates (see Recipes)

3.4 Equipment

1. Thermocycler (Bio-Rad Laboratories, catalog number: 1861096)
2. Light source for visualization of DNA stain (Thermo Fisher Scientific, Invitrogen^{TM}, catalog number: G6600)
3. 12-channel pipette (Mettler Toledo, catalog number: 17013810)
4. Standard gel electrophoresis tank and accessories (Bio-Rad Laboratories, catalog number: 1640302)
5. Autoclave

3.5 Software

1. Geneious v8.0[96] or higher, to design gRNA and repair template (replacement gene). Other gRNA design software can be used as well, such as E-CRISP[101]
2. BLAT[102]

3.6 Procedure

3.6.1 Preparation of CRISPR plasmid

For a diagrammatic overview of the cloning process, see figure 3.1.

1. Design two guide RNA (gRNA) sequences targeting the open reading frame (ORF) for the yeast gene to be replaced using Geneious, or a similar tool such as E-CRISP[101].
   (a) gRNA sequences can often have low activity in practice, despite being predicted to be highly efficient by software tools. In order to minimize setbacks due to a gRNA which turns out to function poorly, we advise designing multiple gRNAs from the outset, and taking them through the cloning steps in parallel, up to and including the construction of the CRISPR plasmids. Both plasmids should then be tested for their ability to target the yeast genome and kill cells (described in later steps) to empirically determine and confirm their activity.
   (b) We have not noticed a strong effect of the location of the gRNA within the ORF. During homologous repair, DNA can be resected up to several kilobases from the break site [103, 104], so the gRNA need not be very close to either terminus of the ORF. It is however important to select a gRNA such that the target site is not present after replacement (i.e. the gRNA should target the yeast ORF, but not the replacement gene).
   (c) Example: For targeting HEM2, the sequences `GGATTATCGGAGATGAA`

`TAG` ("sg1", on the non-coding strand) and `CCTGGTACCAAGGATCCAGT` ("sg2", on the coding strand) were predicted to have high activity (see 3.2).

2. Order forward and reverse oligonucleotides with the gRNA sequence and Golden Gate compatible overlaps:
   (a) Forward oligo consists of the 5 insert `GACTTT` followed by the 20 bp guide sequence specific to the target gene. Example forward oligo for HEM2 sg1 (underline indicates 5 Golden Gate overhang): `GACT TTGGATTATCGGAGATGAATAG`.
   (b) Reverse oligo consists of the 3 insert `AAAC`, followed by the reverse complement of the 20 bp guide sequence, followed by `AA`, which complements part of the `GACTTT` insert on the forward oligo. Example reverse oligo for HEM2 sg1 (underline indicates 3 Golden Gate overhang): `AAACCTATTCATCTCCGATAATCCAA`.
3. Mix forward and reverse oligos (50 µM each) for each gRNA in a total volume of 20 µl and anneal with each other using a thermocycler with the program below. It is unnecessary to phosphorylate the insert.

   95 °C for 5 min

   55 °C for 15 min

   25 °C for 15 min
4. First Golden Gate cloning reaction to transfer into shuttle vector: Set up cloning reaction with annealed oligos and pYTK050 (Table 3.1).

   A 2:1 molar ratio of insert:plasmid is recommended for optimal

Golden Gate cloning of linear DNA.

5. Transform the reaction into competent bacteria and plate with chloramphenicol selection (170 µg/ml). View colonies under UV light and pick the white colonies (those not showing GFP fluorescence), then grow in liquid culture and purify plasmid. The vectors used in Golden Gate reactions described in this protocol are all GFP-dropout vectors: They contain a GFP gene which will be silenced upon successful cloning. Therefore, GFP fluorescence indicates an invalid construct, while successful constructs will lose the GFP gene and the resulting colonies will be white.

   Optionally, the plasmid can be sequenced to check for errors or mutations in the gRNA sequence, such as may occur during synthesis.

6. Second Golden Gate cloning reaction to create gRNA cassette plasmid: Set up cloning reaction which includes connector plasmids ConL1 and ConRE (Table 3.2).

   For best efficiency, all plasmids should be present at the same molarity in plasmid-based Golden Gate assemblies.

7. Transform the reaction into competent bacteria and plate with ampicillin selection (60 µg/ml). View colonies under UV light and pick the white colonies (those not showing GFP fluorescence), then grow in liquid culture and purify plasmid.

8. Third and final Golden Gate cloning reaction to construct the yeast-compatible, complete CRISPR plasmid: Set up Golden Gate

cloning reaction with connector plasmid from the previous step, and yeast Ura backbone plasmid, and Cas9 plasmid (Table 3.3).

9. Transform the reaction into competent bacteria and plate with kanamycin selection (50 µg/ml). View colonies under UV light and pick the white colonies (those not showing GFP fluorescence), then grow in liquid culture and purify plasmid.

The resulting construct is a self-contained CRISPR plasmid, which when transformed into yeast will cause double-stranded breaks (DSBs) at the locus determined by the gRNA sequence cloned into it. 500 ng of this will be used for each yeast transformation, so if multiple replacements are planned, it is helpful to dilute the CRISPR plasmid to a standardized concentration for easier transformation set up later on.

3.6.2 Preparation of repair template DNA

1. Design the template DNA using Geneious or any other cloning software. Obtain the genomic sequence of the target yeast gene ("old gene"), and the coding sequence (CDS) of the replacing gene ("new gene"). The CDS should not contain introns. Create a gene model for the replaced locus by editing the sequence of the old gene so that it contains the new gene in the correct position (i.e., the desired outcome of replacement).

   We find that replacement works best if the original yeast stop codon is left intact. Otherwise, modifying the new gene, for instance to codon optimize for yeast, has proven unnecessary.

2. Design template PCR primers which anneal to about 25 bp of the 5' and 3' ends of the new genes CDS, and also the 5' and 3' UTR immediately adjacent to the ORF (the homology arms). Figure 3.3 shows an example of primer design for replacing the yeast HEM2 gene with its human ortholog ALAD. This process is much easier using the gene model constructed in the previous step: The sequence covering the junction points between yeast genome and the new gene CDS can be used directly as primer sequence.
   (a) The length of the region complementary to the new gene CDS is determined only by standard PCR efficiency concerns, such as melting temperature. This area will serve as a toehold for the first few cycles of the PCR.
   (b) The length of the homology arms is critical for efficient replacement. We find that homologies of at least 70 bp are necessary (in which case the entire primer oligo will be about 90 bp long), and for some genes, 170 bp homologies may be necessary. For even more difficult replacements, longer homology arms can be cloned separately, but we have found that homologies longer than 500 bp are unlikely to increase efficiency further.
3. Use template PCR primers to amplify a large amount of repair template DNA using a high-fidelity polymerase.
   (a) We find that it is helpful to first conduct several test PCRs with different polymerases. Due to the particular design of the template

primers, this PCR can sometimes run inefficiently or generate unwanted non-specific products. Different polymerases have different characteristics, and often a reaction which fails with one polymerase will run efficiently with another, rendering laborious PCR optimization unnecessary.

(b) At least 5 µg of template DNA is needed per yeast transformation, which can usually be obtained from a single 50 µl PCR. Difficult replacements can often be facilitated by using more (10 µg) template DNA, and if multiple transformations are to be performed the amount will also need to be scaled up accordingly. Often several PCRs are necessary to produce enough DNA.

(c) If very large amounts of template DNA are needed, or an efficient PCR is difficult to set up, an alternative method is to clone the template sequence onto a plasmid, which can be amplified in bacteria with the template DNA excised using restriction enzymes.

4. Check the template PCR with agarose gel electrophoresis.

As long as a sufficient amount of the correct template is produced, non-specific products do not necessarily constitute a problem for the replacement. Because the non-specific products usually lack appropriate homologies, they will not be efficiently integrated into the yeast genome. However, if significant amounts of them are present, they will cause over-estimation of template DNA during spectrophotometry-based quantification; thus the amount of template DNA used in the transformation

would need to be adjusted accordingly. Alternatively, the PCR can be optimized to reduce non-specific products, or only the correct product can be quantified from the gel using a DNA ladder calibrated for quantity estimation.

5. Purify template PCR using the Zymo DNA Clean&Concentrator-25 kit. Elute in double distilled water.

Ideally, the volume of DNA included in yeast transformation should be small, so as to not interfere with the transformation reagents. The elution volume should be adjusted accordingly so that the resulting concentration of DNA is not too low. In our experiments, we have found that eluting with 25 µl double distilled water will usually yield 400-800 ng/µl DNA, which is suitable for transformations.

3.6.3 Yeast transformation

1. Prepare competent yeast cells using the Zymo EZ competent yeast kit according to the kit instructions.

   The EZ 1 solution in this kit can be substituted with 100 mM lithium acetate without significant change in transformation efficiency.

   The amounts given in the kit manual can be slightly modified: 2 ml yeast culture can be used to produce 100 µl of competent yeast, which is sufficient for two transformations, 50 µl each.

2. Set up a transformation reaction: Mix 50 µl competent yeast, 500 µl EZ 3 solution, 500 ng of CRISPR plasmid and 5 µg repair template DNA

(up to 50 µl total volume). Incubate at 30 °C as directed by kit manual and plate on Ura medium.

When using a new gRNA for the first time, gRNA efficiency can be estimated with a control transformation, which is performed as stated but without repair DNA. When the CRISPR plasmid is introduced without a repair template, it will repeatedly cleave the target locus, causing toxicity. Very few or no colonies are the ideal outcome, since this indicates highly efficient CRISPR cleavage and low background rate. Cells can survive the CRISPR plasmid uptake without repair DNA if the CRISPR activity is stochastically low (such as due to poor gRNA efficiency) or mutations at the CRISPR target locus can be tolerated (which produces false transformants even in presence of the repair template).

3. When colonies appear on the Ura plates, collect up to 12 of them with a pipette tip and suspend in 50 µl water. These suspensions will be screened for confirmed replacements. Yeast suspensions can be stored at 4 °C and used to start new cultures for up to 2 weeks.
   (a) Typically, colonies will appear on Ura plates (Figure 3.4) after 1-3 days. In some cases, the replacement will impose a significant fitness defect such that up to 6 days may be required for colonies to appear, but we have not encountered cases where colonies from a successful transformation take longer than 6 days to grow.
   (b) The uracil dropout medium will select against cells which failed to take up the CRISPR plasmid (which confers uracil prototrophy),

but because the CRISPR plasmid is toxic to cells unless a successful replacement occurs (eliminating the CRISPR target locus) only cells which have a replaced locus are expected to survive. However, due to spontaneous hypoactivity of the CRISPR system, mutations in the CRISPR target locus[20], and cells which manage to survive CRISPR-associated DSBs, there will be a background rate in the form of false transformant colonies which do not carry the correct genomic replacements. To save time, we recommend collecting several transformant colonies and screening them in parallel.

(c) To streamline this process (especially when several replacements are performed in parallel), pick colonies with pipette tips and manually attach them to a multichannel pipette (Figure 3.5). The multichannel pipette can then be used to suspend all 12 samples in one row of small PCR tubes or a 96-well plate.

3.6.4 Colony screening via PCR

1. Design confirmation PCR primers: Primer pairs should be selected such that the forward primer anneals to the yeast UTR while the reverse primer anneals only to the new gene CDS but not the old gene's ORF. Thus, the product should span the junction point between foreign sequence and native yeast genome. The yeast UTR primer should preferably not overlap the homology region.
   (a) Ideally, the product size should be small, about 300 bp, for a faster

and more robust PCR.

   (b) It is sufficient to check only the 5' junction point, since it is rare for integration to proceed as expected at one end of the gene but introduce artifacts at the other.

   (c) If desired, the absence of the yeast ORF can also be tested by using a reverse primer which anneals to yeast ORF only. However, lack of product from such a primer pair is not sufficient to confirm a clone, since the reaction is liable to fail for unrelated reasons (such as poor lysis of cells).

2. Prepare lysates of harvested transformants: Mix 5 µl of each yeast suspension with 15 µl zymolyase solution.

3. Incubate lysates for 30 min at room temperature, then 15 min at 37 °C and 5 min at 95 °C.

4. Set up 20 µl colony PCRs with confirmation primers and using Accuprime Pfx as the polymerase. Use 1 µl of the lysate as template DNA.

   (a) We find that other polymerases do not perform well due to impurities from the yeast lysates. Due to the impurities introduced by the lysate, the colony PCR may spontaneously fail, leading to false negatives. To ameliorate this problem, a positive control PCR can be performed for each lysate, which is identical to the confirmation PCR but uses primers complementary to an unrelated, unmodified locus in the genome. We use two primers targeting a 500 bp segment of the yeast ERG13 promoter for this purpose (forward `CGA`

`ACTGGATGAGATGGCCG` and reverse `CATGCTGCACCTTTTATAGTAATTTGGC`).

5. Check the colony PCRs for product by agarose electrophoresis. Lysates from clones with the correct modifications should generate a product with the confirmation primers. Background false transformants (e.g. mutants) will not produce a band.
   (a) A PCR product from the confirmation primers is sufficient evidence of successful integration of the repair template. For further verification, the locus can be sequenced, but we have found that dramatic sequence artifacts rarely occur in clones confirmed by PCR, the most common mutations are single-basepair substitutions or indels, which typically constitute a minority of confirmed clones.
   (b) Lack of product from the confirmation primers is inconclusive per se. In such cases, it is worthwhile to consider additional evidence, such as whether the positive control PCR worked (if not, the lysis may have failed).
6. Confirmed clones can be propagated by starting a new culture from the original suspensions of yeast in water.

3.6.5 Curing of the CRISPR plasmid

1. Streak original water suspensions of confirmed clones on YPD. The CRISPR plasmid is low copy and can be spontaneously lost in absence of selection.

2. Pick 10 colonies from the YPD plate and patch each one on YPD and SD-Ura plates.
3. Incubate both plates, and collect cells from patches which grew only on YPD but not on SD-Ura. Isolates which still carry the CRISPR plasmid will grow on uracil dropout medium, but those which have lost the plasmid will not. Typically, 3 days is sufficient to confirm lack of Ura prototrophy, but if slow growth on uracil dropout is suspected, incubation can be extended to up to 6 days to definitively confirm no growth on uracil dropout.

   The plasmid can also be cured by counterselecting on 5-fluoroorotic acid (FOA) plates[105]. However, there is a possibility that this FOA method will generate some colonies that are not cured of the plasmid but rather have acquired a mutation in the Ura marker (thus continuing to express the gRNA). Thus, FOA counterselection should not be used (as opposed to replicate patches on YPD and Ura) if it is important to ensure curing of the plasmid, rather than simply abrogating Ura prototrophy. On the other hand, the FOA method can save time if only loss of Ura heterotrophy is desired, for instance to enable a subsequent transformation with a different Ura-selectable plasmid.

3.7 Data Analysis

The data analysis needs for this procedure are minimal. Most importantly, when using Geneious to design gRNA sequences, it is desirable to

select gRNA sequences that have high predicted on-target activity (automatically calculated by Geneious). gRNA sequences with high predicted activity may have low actual activity, but they will be less likely to exhibit low activity than sequences with low predicted activity. The distance of the gRNA target site can be up to 1 kb away from either homology region without perceptible negative consequence, thus gRNAs should be selected primarily based on high activity rather than location (provided that they lie between the two homology arms).

3.8 Notes

1. We have found that even among gRNAs with high predicted activity, some will fail to induce double-strand breaks with sufficient efficiency for editing. It is highly recommended that for each target locus, several gRNA are designed and tested in parallel, to ensure that at least one will be a sufficiently good DSB inducer for purposes of genome editing.
2. If a given gRNA exhibits significant off-target activity, the likely outcome is that off-target cleavage will kill most of the transformed yeast cells. Successful, efficient genome editing in yeast relies on lethality associated with DSBs at the target locus being rescued by HR (allowing efficient repair of the DSB) and abrogation of the gRNA target site (preventing further cleavage). In the event off-target activity, HR may likely not take place because no repair template with homology to the off-target site has been supplied, moreover the gRNA site will not be eliminated for the same reason. Further, the confirmation strategy we suggest is such that only repair at the correct locus will produce a positive result. However, it is nevertheless worthwhile to ensure that selected gRNA target sites do not occur at other locations in the genome, where cleavage is not intended. Although it is very unlikely for the combined 23 bp target sequence to appear multiple times in the yeast genome, we recommend confirming that candidate gRNA sites appear only in the target locus using a tool such as BLAT.
3. gRNA targets consist of a 20 bp sequence (which will also be included

in sgRNA sequence and become part of the Cas9 complex) followed by a 3 bp PAM sequence (which takes the form of `NGG` for Cas9 described in this protocol). The PAM sequence does not become part of the gRNA, but it must be present in the target genome for Cas9 cleavage to occur. This can be verified by attempting to align the gRNA sequence to the sequence of the repair template typically, CRISPR activity will be very low with more than 5 mismatching basepairs, although mismatches in the PAM and proximal to the PAM appear to have more significance [106]. When replacing with very similar sequences, such that it is difficult to find good gRNA sites unique to the target locus, one strategy that can be adopted is to introduce synonymous mutations in the repair template sequence which alter the PAM site or PAM-proximal nucleotides. Alternatively, recent research suggests that using shorter gRNA may increase specificity, since the 8-17 PAM-proximal nucleotides contribute disproportionately to CRISPR target recognition[107].

4. There is some variability in the yeast transformation step, and depending on how the competent cells were prepared, and how the transformation was performed. Most commonly, the number of resulting colonies will vary somewhat between transformations of identical strains with identical reagents, but usually this variation will be less than tenfold. When a transformation produces a fair number of colonies (at least 10) yet none of them are found to be correct clones upon screening, simply repeating the transformation is unlikely to improve results. The most straightfor-

ward avenues of increasing the number of correct clones are to increase the amount of repair template DNA, and to produce repair template DNA with longer homologies.

5. If no colonies appear after transformation, the reason may be low transformation efficiency. In this case, several troubleshooting steps can be taken (described in detail in the documentation of the Zymo EZ competent yeast kit). We have found the following to be effective:
   (a) Thoroughly vortexing the mixture of competent cells and DNA.
   (b) Longer incubation time for the transformation (1.5 h instead of the 45 min).
   (c) Including more cells in the transformation.
   (d) Competent cells seem to perform slightly better when frozen once (slowly in -80 °C) than freshly prepared cells.
6. When the CRISPR reagents and repair template are transformed into yeast cells, the resulting transforming colonies will be of three kinds with respect to the targeted locus:
   (a) Correct transformants which bear the sequence of the repair template.
   (b) False transformants which bear the original, unedited sequence.
   (c) Mutants.
7. In our experiments, we have found that the first two classes predominate unless mutants are specifically selected for. Even in the absence of a repair template, the majority of false transformants will not be mutants.

Due to the efficient HR system of *S. cerevisiae*, if the conditions of the experiment are adequate then editing will take place at a very high rate. Thus, typically, the proportion between the first two of the three classes listed above will be such that the transformants are either mostly correct or all false. The third class, or mutants, we have found to be very rare in either case unless specifically selected for. As a consequence, it is rarely necessary to screen a very large number of colonies to determine whether an editing experiment has succeeded. However, it is desirable to collect several confirmed clones to minimize issues caused by artifacts, such as mutant edited sequence caused by errors during PCR (with the reagents and protocols described in this text, we have found clones with mutant edited sequence also be very rare).

8. Selecting yeast transformants with a single amino-acid dropout medium is normally a straightforward process, and colonies can be seen within 1-2 days of plating. However, occasionally the genome editing process itself, or the resulting edited sequence, can result in a growth defect in the resulting cells. Thus, if no colonies appear, incubating the plate for a longer period can produce colonies. In the most extreme case we observed, it took 6 days for colonies to appear on a uracil dropout medium, but several clones were later confirmed by PCR and sequencing; these clones consistently exhibited slow growth in subsequent culture on rich medium (YPD) as well.
9. Some combinations of target locus and repair template may lead to a

mixture of large and small yeast colonies after transformation. If this occurs, generally it is best to screen an adequate number of colonies for each size class. It may be that the correct edits create much slower growing strains, thus the large colonies are false while the small ones have the desired edit. Conversely, if the desired sequence does not interfere with normal growth, but mutations arising from NHEJ do, then larger colonies will tend to be the correct clones. We have observed examples of either case when humanizing and bacterializing various loci. It is difficult to predict a priori which case will be evident for a given transformation, therefore it is often more practical to screen colonies and recording their size, and also ensuring that each size is adequately represented in the screen.

10. When picking colonies for the colony PCR screen, only a small quantity of cells is needed. Most likely as little as 1,000 cells will be sufficient to obtain a PCR product. We have often chosen to collect slightly larger numbers of cells to visually confirm their suspension in water by turbidity. However, too many cells lead to incomplete lysis and inhibition of the colony PCR. With cell clumps larger than 1-2 mm the colony PCR will often fail. So ideally, the cells collected from the colony should form only a tiny speck, 0.5 mm or smaller in diameter. It is helpful to include the positive control PCR when screening, to identify samples which failed to produce a PCR product due to poor lysis. Lysis and PCR can be repeated for these samples if needed.

11. It is possible to adapt the protocol described here for the simultaneous replacement of multiple genes. The Mo Clo toolkit allows for cloning up to 4 different gRNA cassettes on the same CRISPR plasmid; for this, the gRNAs would be captured on pYTK050 as described here, but in the second Golden Gate reaction, instead of the ConL1 and ConRE plasmids, the first gRNA would be cloned with ConL1 and ConR2, the second with ConL2 and ConR3, the third with ConL3 and ConR4 and the fourth with ConL4 and ConRE (this process is explained in detail in [23]). All of these cassette plasmids would then be included in the final Golden Gate reaction to assemble the CRISPR plasmid. Then, during transformation of yeast, templates for each of the included gRNAs will need to be co-transformed. However, multiple replacements are even more dependent on efficient transformation, cleavage and repair than single replacements, and some additional work may be necessary to optimize these parameters in practice.

3.9 Recipes

1. Zymolyase solution (50 ml)
   (a) Weigh 9.11 g D-sorbitol
   (b) Dissolve in 50 ml distilled, deionized water to make 1 M sorbitol and autoclave
   (c) Weigh 0.25 g zymolyase and dissolve in sorbitol solution
   (d) Aliquot and store at -20 °C
2. Lithium acetate, 100 mM (40 ml)
   (a) Weigh 0.408 g lithium acetate dehydrate
   (b) Dissolve in 40 ml distilled, deionized water
   (c) Filter sterilize (0.2 µm filter) and store at room temperature
3. LB medium (1 L)
   (a) Weigh 25 g LB powder
   (b) For solid medium, add 15 g agar
   (c) Dissolve in distilled, deionized water for 1 L total volume
   (d) Autoclave and let it cool to 60-70 °C
   (e) Pour in Petri plates so that the medium covers the visible area of the plate
   (f) Let plates cool and solidify at room temperature, store at 4 °C
4. YPD (1 L)
   (a) Weigh 50 g YPD powder
   (b) For solid medium, add 20 g agar
   (c) Dissolve in distilled, deionized water for 1 L total volume

(d) Autoclave and let it cool to 60-70 °C

(e) Pour in Petri plates so that the medium covers the visible area of the plate

(f) Let plates cool and solidify at room temperature, store at 4 °C

5. SD-Ura (1 L)

(a) Weigh 1.5 g yeast nitrogen base w/o amino acids, 5 g ammonium sulfate, 20 g dextrose, 2 g SC-Ura dropout powder

(b) For solid medium, add 20 g agar

(c) Dissolve in distilled, deionized water for 1 L total volume

(d) Autoclave and let it cool to 60-70 °C

(e) Pour in Petri plates so that the medium covers the visible area of the plate

(f) Let plates cool and solidify at room temperature, store at 4 °C

3.10 Acknowledgments

This work was supported by grants from the NIH (R21 GM119021, R01 HD085901, DP1 GM106408, R01 DK110520, R35 GM122480), Army Research Office (ARO) grant W911NF-1210390, and the Welch Foundation (F-1515) to E.M.M. We would like to thank John Dueber & colleagues for producing the excellent plasmid toolkit which greatly facilitated our work.

3.11 Conclusion

This work was originally published in the journal *Bio-protocol*. I contributed to much of the development of the technique and the writing of the manuscript. Soon after publication, we were serendipitously informed that another group had succeeded in reproducing our results.

Though the protocol is written with a focus on ortholog swapping, it is in principle equally well suited to any arbitrary genetic editing. In addition, due to the inherent DNA cleavage activity of CRISPR, the strains produced by this protocol can be used as gene-drives[108] with super-Mendelian inheritance of the resulting edit in subsequent generations (Figure 3.6). Ordinarily, if a haploid yeast *a::B* (with sequence B replacing wild-type locus A) mates with a wild-type yeast the result will be a heterozygous A locus with AB genotype. If the CRISPR plasmid is retained (by skipping the plasmid curing step at the end of the protocol), then after mating the CRISPR system will cleave the wild-type A allele coming from the unedited parent. Once a double strand break occurs, the edited chromosome (which is immune to CRISPR cleavage by

virtue of lacking the gRNA target site) will be used as a template for HR. The practical result is that haploid parents of genotypes *A* and *a::B* are mated, but instead of *AB* progeny, *BB* progeny is the result. Thus the practical function of a CRISPR plasmid during mating is to eliminate homozygosity at a certain locus, preferentially with respect to the wild-type allele. If such a plasmid (often referred to as a gene drive) is maintained throughout generations, the allele favored by it will propagate through the population rapidly and eliminate alternative alleles. Gene-drives are therefore a powerful tool for combining a set of different edits or propagating a single edit throughout a large population rapidly.

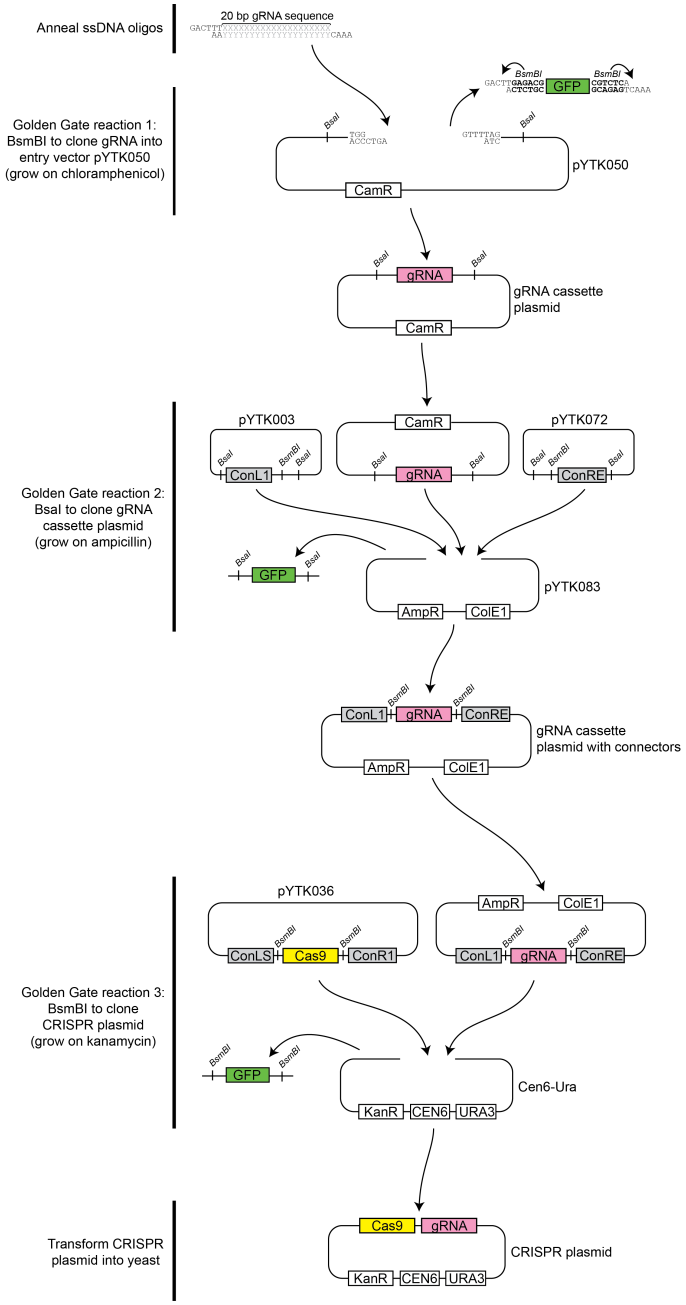

Figure 3.1: Overview of the CRISPR plasmid construction process. In the first step Xs and Ys represent the gRNA sequence selected, and BsmBI recognition site is indicated in bold.

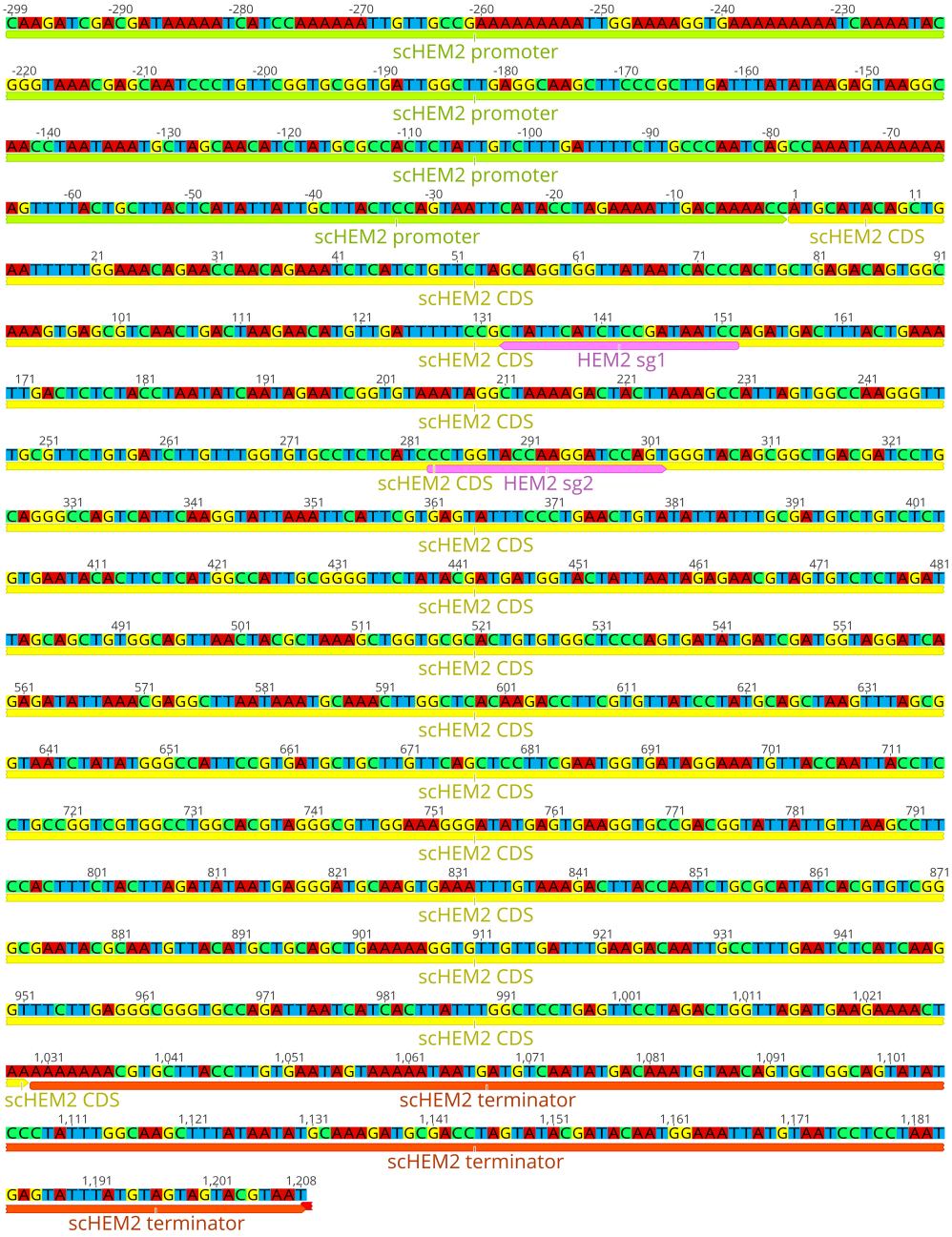

Figure 3.2: Diagram of the native yeast HEM2 locus showing positions of the example guide RNAs sg1 and sg2.

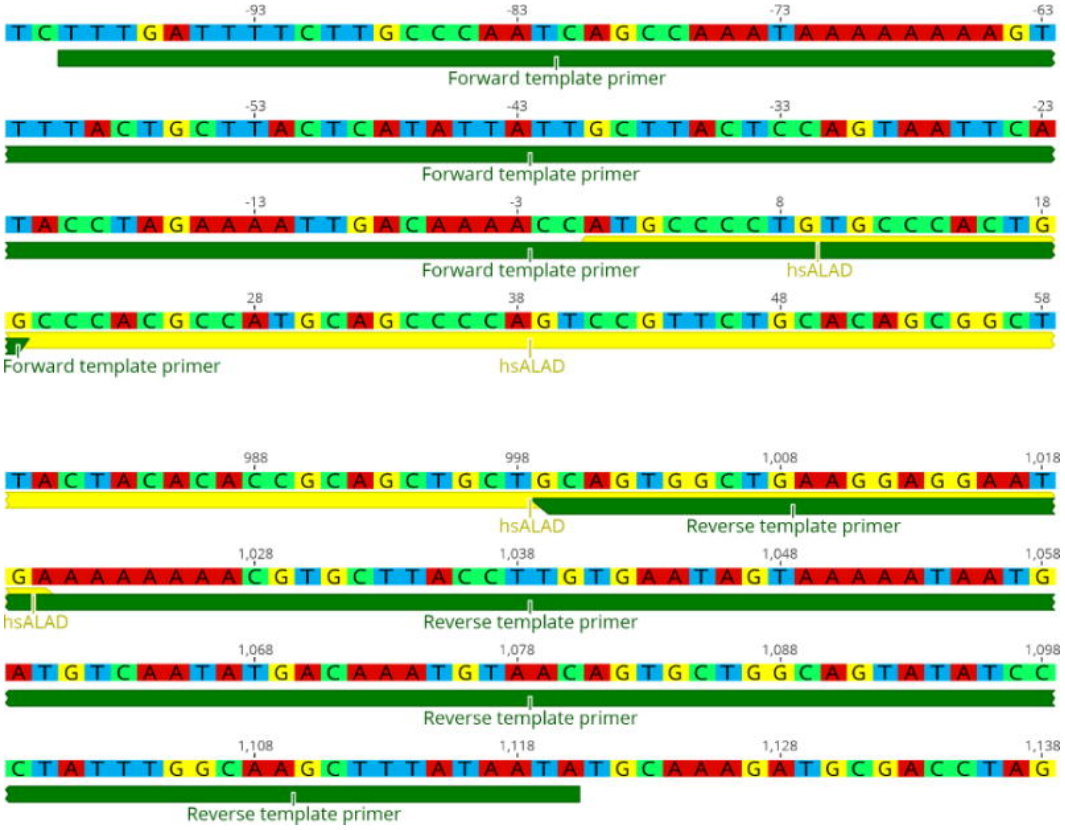

Figure 3.3: Diagrams of example template primer designs for the replacement of HEM2 with hsALAD.

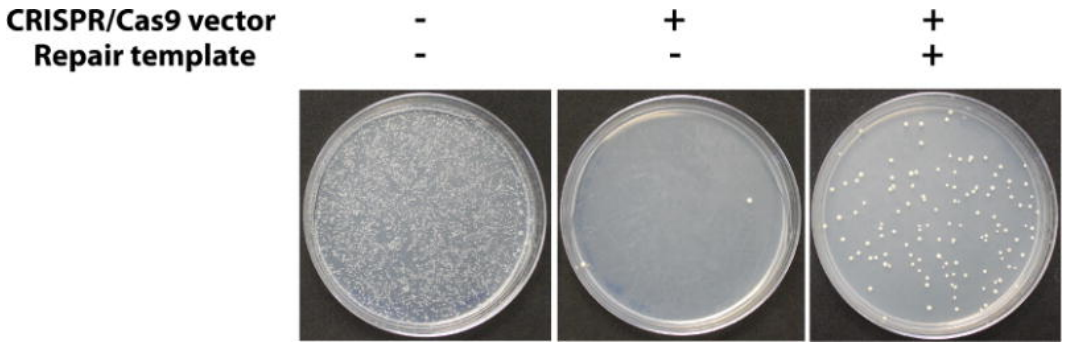

Figure 3.4: Representative assay results. Yeast cells are rescued from DSB lethality (center plate) when an appropriate repair template is provided (right plate). The left plate is a negative control of cells carrying a control plasmid with the same selectable marker (URA3) done to estimate the transformation efficiency of the yeast strains being used.

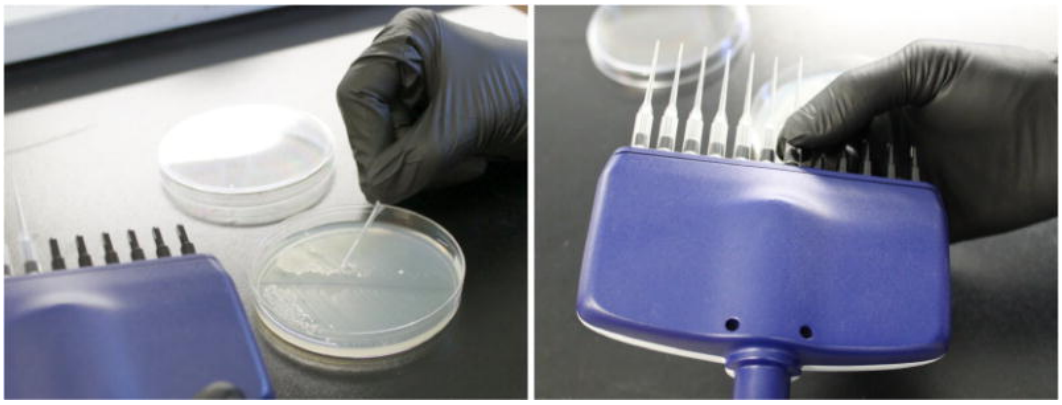

Figure 3.5: Demonstration of colony picking technique with 12-channel pipette.

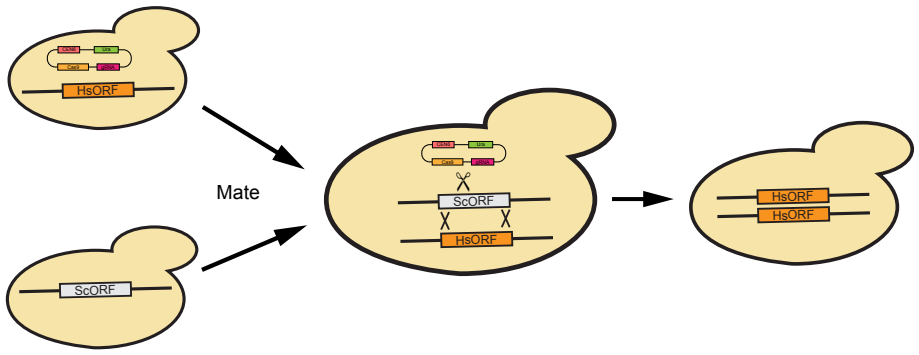

Figure 3.6: CRISPR/Cas9 as a gene drive. Ordinarily, CRISPR-mediated humanization is carried out in haploid strains. A humanized strain is mated with another yeast of the opposite mating type, but not humanized at that locus. This generates a heterozygote (yeast/human) diploid. If the CRISPR plasmid targeting this locus is retained during the mating, it will also cleave the wild type allele received from the other yeast. The homologous chromosome can serve as a repair template, and the yeast allele is overwritten with the human one during homologous recombination. If the plasmid was not preserved, or it is desirable to decouple the HR from the mating itself, it can still be transformed later with the same effect.

Reagent	**Amount**
dsOligo	40 fmol
pYTK050	20 fmol
NEB T4 buffer 10x	1.0 µl
NEB T7 ligase	0.5 µl
NEB BsmBI	0.5 µl
H_2O	to 10 µl

Table 3.1: Golden Gate reaction for cloning into shuttle vector.

Reagent	**Amount**
gRNA on pYTK050	20 fmol
ConL1 (pYTK003)	20 fmol
ConRE (pYTK072)	20 fmol
AmpR-ColE1 (pYTK083)	20 fmol
NEB T4 buffer 10x	1.0 µl
NEB T7 ligase	0.5 µl
NEB BsaI	0.5 µl
H_2O	to 10 µl

Table 3.2: Golden Gate reaction for cloning gRNA cassette plasmid.

Reagent	**Amount**
gRNA cassette plasmid with connectors	20 fmol
Cen6-Ura cassette*	20 fmol
Cas9 plasmid (pYTK036)	20 fmol
NEB T4 buffer 10x	1.0 µl
NEB T7 ligase	0.5 µl
NEB BsmBI	0.5 µl
H_2O	to 10 µl

Table 3.3: Golden Gate reaction for cloning CRISPR plasmid.

*Cen6-Ura is constructed by assembling YTK plasmids (008, 047, 073, 074, 081, and 084).

Chapter 4

Multiple Humanization of the Heme Pathway

4.1 Introduction

In past work, we observed that modularity largely determines swappability. I have also noticed that although many genes can be swapped, many of these replacements cause slight fitness defects. Motivated by these findings, I questioned the limits of swappability: Is it possible to replace multiple genes within the same yeast? What are the consequences of doing so? This chapter describes my work on investigating these questions, primarily in the context of humanizing the yeast heme biosynthesis pathway.

Detailed consideration reveals several interesting aspects to this question, and describes the rationale and motivation for attempting to humanize the heme pathway of the yeast.

4.1.1 Is multiple humanization feasible?

One of the simplest naive expectations is that if many genes can be individually humanized, then they can be humanized together as well. This corresponds to the reductionist null hypothesis that replacements do not significantly influence each other. It is instructive to consider a similar case in the recent history of yeast biology: The striking discovery that most yeast genes are non-essential[109, 6] was soon followed by the equally striking discovery that there in many pairs of such dispensable genes are synthetic-lethal[12, 110], and become essential once one of them is deleted. In fact, the so-called non-essential yeast genes are so considered because they are non-essential in culture on rich medium under standard laboratory conditions. For most es-

sential genes, there exists at least one environment where they do become essential[7]. It appears that often, genes can appear to be dispensable because their function is redundant, but when the redundancy is eliminated through environmental stress or genetic disruption, they cease to be dispensable.

Analogously, we hypothesize that many cases of yeast humanizations are not necessarily complete complementations of the native gene's function. Humanizing a gene only shows that the *non-redundant* functions of that gene have been provided by the human ortholog. It may be that the human ortholog lacks key functions that the yeast gene has, or has additional functions that, in a yeast cell, would be deleterious. In single humanizations, these functional impairments could be overcome due to other native yeast genes acting as a redundancy for the same functions. Higher order humanization can conceivably eliminate all redundant genetic providers of an essential activity, thereby producing a "surprising" behavior of mutual exclusivity between humanizations which are possible individual. Constructing a multiply humanized yeast helps elucidate this hypothesis and a number of long-standing theoretical concerns. Moreover, focusing on a single pathway increases the relevance of such an experiment, as we would expect key metabolic modules to be particularly rich in epistasis.

4.1.2 Does compatibility extend to pathways?

Past studies revealed modularity plays a large role in which yeast genes are amenable to ortholog swapping. Genes which can be readily humanized

often belong to pathways or protein complexes (collectively, the "modules") made up of other swappable genes. This was a major discovery of the initial systematic humanization project[8]. It was also confirmed in our later work[37] generalize across domains and kingdoms of life (Figure 2.1). It is tempting to rationalize this phenomenon in terms of compatibility within the module: We can expect that evolutionary divergence would introduce distinct molecular idiosyncrasies to orthologous genetic modules[111, 112]. This would hamper cross-species swapping, not necessarily due to a deficiency in the function of the ortholog *per se*, but due to its inability to act in concert with the whole.

In the course of applying genomic editing to humanization research, it became apparent that often humanization can lead to a modest fitness defect. In older studies, human orthologs were often provided on plasmids, driven by strong constitutive promoters. This likely expressed many human genes at a much higher level than their native counterparts, and also blocked many native regulation mechanisms. With the CRISPR-based genomic replacement approach described in Chapter 3, the orthologous ORF directly takes the place of the yeast ORF and is amenable to many of the same mechanisms of transcriptional regulation. It is, therefore, possible to directly measure the fitness defects associated with the swap.

If each genetic modification i results in fitness f_i (relative to the wild type) when performed individually, and if they are independent of each other, we could expect that combining the modifications will result in fitness $F = \prod_i f_i$ (the "multiplicative model"). We could expect this behavior if the fitness de-

fect of humanization is determined by interactions with the entirety of the genome or the hypofunction of the human ortholog itself. However, if interactions with other module member are more important, they will not be disrupted by bringing an entire human module into yeast wholesale. When humanizing the last gene in a yeast module, we would predict little fitness defect as the human interaction partners of that gene would already be present. Generalizing from this by induction, we can predict that fitness would depend on how much of the module has already been humanized. The initial humanization steps would gradually impair growth, but after the lion's share of the module is human, there would be a tipping point and subsequent humanizations would *improve* fitness – for the same reason that "yeastizing" a human cell might decrease it. Recalling our earlier formalism, the fitness cost of humanizing each gene f_i would depend on fraction r of the module that has already humanized (the "cooperative model"). An illustration of these models is provided in Figure 4.1.

Humanizing an entire yeast pathway allows us to directly test these alternative explanations. Moreover, there is also a practical benefit: The multiplicative model implies that it is possible to humanize only so much of a yeast genome. This constrains the suitability of using humanized yeast clinically, as a test bed for investigating the biology of human genetic modules, because it predicts that any substantial human pathway will cause the strain carrying it to be extremely slow-growing. As a result, strategies would need to be developed to cull large modules to a handful of key genes in order to balance

fidelity to human physiology with experimental practicalities. Conversely, if the cooperative model is true, then the solution to slow growth of a humanized yeast system is paradoxically to humanize it more rather than less.

4.1.3 Scalable probing of epistasis and allelic variation

The CRISPR-based genome editing system I am utilizing for this research has a convenient side-benefit: After a humanization is finished, the CRISPR plasmid can act as a gene drive to greatly simplify combining individual humanizations (Chapter 3, section Conclusion). When introducing a new humanization, it is necessary to transform a linear DNA that serves as the repair template. The success of the procedure, therefore, hinges not just on the ability of the human ortholog to complement its function, but also on efficient transformation and incorporation of the template. In contrast, when inducing loss-of-heterozygosity via a CRISPR gene drive, efficiency of the genome editing is very high: The entire chromosomes act as homology arms, and transformation efficiency is no longer a factor. This permits reliable humanization with not just wild type human alleles, but also allelic combination with significant deleterious epistasis.

Under the cooperative model given above, the number of humanized strains that can be constructed for a module with n genes is given by $\sum_{i=0}^{n} \binom{n}{i}$. For 8 genes (as in the heme pathway), this evaluates to 255. It would be impractical to construct so many strains individually, but the gene drive permits a high throughput approach, analogous to existing techniques based on large

yeast crosses[11, 113]. Conducting an all-by-all cross of partially humanized strains affords deep knowledge of the epistasis network inherent in the clinically relevant heme pathway. Moreover, creating this system would enable its generalization to also probe various allelic variants of human genes and provide a very practical, scalable, economical method of probing massive human variation data.

4.2 Higher order humanization of the yeast heme biosynthesis pathway

I began efforts at multiple humanization with the collection of individually humanized strains generated as part of the work described in [37] and elsewhere. In transformation experiments, UROS, UROD, and HMBS seemed to be the human genes that can be humanized most efficiently. In separate experiments, UROS was transformed with HMBS and UROD. UROD was transformed with UROS and HMBS. HMBS was transformed with UROS and UROD. The transformation was as described in Chapter 3. I was able to confirm a large number of UROS/UROD transformant colonies, which were afterward dubbed the "UU" strains. All singly humanized strains exhibited a small but perceptible slowness in growth, which was evident from colony size when they were cultured in parallel with wild type yeast. The UU strain appeared to have a similar slight fitness defect.

The UU strain was cured of the CRISPR plasmid using the 5-FOA counterselection method and, in separate experiments, transformed with ALAD, HMBS, CPOX, and PPOX. Once again I obtained a large number of transformants for ALAD/UROS/UROD ("AUU") and HMBS/UROS/UROD ("HUU") and confirmed them by colony PCR. The CPOX and PPOX experiments produced very few colonies (typically <5) and none were confirmed by PCR, indicating that they likely failed to incorporate the humanized template and repaired CRISPR-induced damage by NHEJ. It has often, though not always, been the case that false transformants in genomic editing experiments

will grow much slower. NHEJ is known to be mutagenic in yeast, with mutation rates on the order of 0.1%[20]. It is possible that repeated cleavage in these failed transformants (although the plasmid bearing their CRISPR system was continuously selected for, the repair template is only transiently present) induced and selected for mutations that eliminated the CRISPR target site, but at the same time impaired the activity of an essential gene. It is also likely that the slow growth is due to repeated CRISPR cleavage which the clones manage to tolerate. In other experiments, Sanger sequencing has shown that many false transformants do not have any mutations in the area around the CRISPR target site.

Both AUU and HUU clones all showed a moderate growth defect, somewhat greater than that of UU and singles. Both clones were cured of plasmid, grown out and prepared for a subsequent round of transformation with PPOX, FECH, HMBS (for AUU) and ALAD (for HUU). The transformation succeeded for ALAD, generating the ALAD/HMBS/UROS/UROD strain ("AUUH"). I isolated several transformant colonies, confirmed them by PCR, and selected one for further experimentation. Interestingly, the HMBS transformation which would have generated the same genotype failed to produce any confirmed clones. Moreover, it resulted in a very small number of transformant colonies (<5) which has often been a hallmark of a failed replacement in our system. A second attempt to transform HMBS into AUU also failed in similar fashion. This somewhat paradoxical result appears to imply that swappability can depend not merely on the set of genes being humanized, but

also on the order in which they are humanized. I suspect that the paradox is explained by the nature of our experimental method: The limiting factor in our technique is transformation efficiency, as even the most successful experiments will produce at most around 50 colonies transformed with the plasmid. Typically, the number is much smaller, in the 10-20 range. There is no direct selection for the repair template, so one would expect a large number of colonies passing -Ura selection may have failed to be co-transformed with the template as well. The lethality of CRISPR activity in absence of HR should kill most such partial transformants, but in practice, some false transformants are often recovered from replacement experiments. Therefore, in many cases, the probability of recovering correct clones is not much greater than the limit of detection. Any genetic change, such as humanization of multiple genes, could impair cellular function just enough for transformation efficiency fall below that limit. Given that transformation itself, as well as successful HR, is not a trivial physiological event, it is not unrealistic to suspect that certain combinations of humanized genes interfere with it.

The AUUH strain proved to be the most extensively humanized strain I was able to construct. Notably, it had an extremely slow growth rate. Typically, Ura prototroph yeast will form visible colonies on -Ura medium after 1-2 days[114]. In accordance with the fitness defect imposed by humanization, colonies from successful transformations will often grow slower, appearing only after 2-4 days. Colonies from the AUUH strain had only reached a large enough size to be picked (2-3 mm in diameter) after 6 days of culture at 30 °C. Al-

though such slow growth is often an indicator that the colonies are false and do not contain the humanization, we were able to definitively confirm the identity of these clones in multiple colony PCR screens (Figure 4.2).

Transformations of AUU and HUU with either PPOX or FECH failed to produce any confirmed clones. This result is unsurprising given the earlier difficulties with humanizing PPOX in UU context. Despite several repeated attempts, I was likewise unable to recover any successful clones from transforming AUUH with FECH, ALAS1, CPOX or PPOX. Interestingly, when I succeeded in obtaining single ALAS1 and FECH strains, the humanization appeared very inefficient, generating only a few clones. The clones that were confirmed grew very slowly, between AUUH and AUU/HUU despite being only singly humanized. Further, the FECH clone showed a strong pink cell phenotype which we had earlier observed for ecolization of the same locus (Figure 2.12)

4.3 Evolutionary optimization of the AUUH strain

In order to facilitate experimental work with the quadruple strain, and also to gain a deeper understanding of the biological basis for its slow growth, I next conducted a suppressor screen to discern the selective consequences of quadruple humanization.

In order to study the feasibility of obtaining suppressor mutants in this way, I first conducted a preliminary solid culture experiment. A clonal, validated stock of AUUH was plated on a large YPD agar plate at a sufficient

density to generate large numbers of well-separated colonies. After 3 days of culture, there was a clear and consistent pattern of bimodal colony sizes (Figure 4.3). To confirm that the large colonies were bona fide suppressors rather than contaminants, I collected 12 with sterile toothpicks for a colony PCR screen. In order to further verify the results and provide a control, I likewise collected 12 smaller colonies, which were presumably also AUUH populations that had not acquired significant suppressors. All 24 colonies in this set were confirmed to steal bear humanizations at all 4 loci. This demonstrated the feasibility of obtaining much faster growing mutants of AUUH by relying only on the background mutation rate of yeast.

I then conducted a more elaborate evolution experiment. A clonal AUUH stock was split into 8 sister lineages, and each was passaged daily in liquid YPD for 12 days. The final sample from each lineage was designated A12-H12, with letters identifying the lineage and numbers indicating the day. All 8 evolved lineages, as well as the ancestral AUUH and BY strains, were then assayed for fitness by measuring growth rate in triplicate4.5. Although there was substantial variation between lineages, all were significantly more fit than the ancestral strain, with the fastest growing strain being C12 4.6. The results of the liquid culture passage experiment corroborate the feasibility study on solid culture: The relatively consistent amount of fitness recovery across distinct cultures, and the significant change over such a short time period, suggests that some "low hanging fruit" must have been available in the evolutionary landscape of the AUUH strain. Likely, all lineages were explor-

ing a similar clique of compensatory mutations, which were also exhibited by the colonies in my feasibility study, and help explain the bimodal colony size distribution.

4.4 Whole genome sequencing

In order to investigate the mutations accrued by my evolved lineages, I performed whole genome sequencing of several evolved lineages, along with the ancestral strain to establish the baseline, and the BY4741 to control for mutations in our lab stock. The reads were mapped to the reference genome of BY4741.

Coverage was evenly distributed between strains and chromosomes, at a mean of about 15x (Figure 4.7). The notable exception was chromosome XII which was covered at 30x in the initial mapping for all strains including BY4741. Chromosome XII contains the highly repetitive ribosomal DNA (rDNA) sequences[115, 116], which accounts for this — there are typically more than a hundred repeats of rDNA and collectively they make up around 60% of chromosome XII. The reference sequence for chromosome XII (`NC_001144.5`, [117]) represents rDNA as a single copy, thus all of the reads from the repeat array were collapsed into this single locus during mapping (Figure 4.8). With reads mapping to rDNA excluded, the chromosome is covered at a mean of 15x, which was the value used in figure 4.7. Coincidentally, chromosome XII does not contain any of the heme loci.

Notably, there were conspicuous gaps in coverage around the cod-

ing sequences of genes HEM2, HEM3, HEM4, HEM12 for AUUH but not BY4741. This confirms the presence of these humanizations in AUUH (Figures 4.9–4.12).

The number of short variations was generally slightly smaller in the ancestral AUUH than in BY. However, the evolved strains had a much higher number of variants, confirming that substantial evolution has taken place in the passage experiment (Figure 4.16). Interestingly, it appeared that for chromosome I, the number of variants was much smaller in AUUH than in BY, but the evolved strains appear to have accumulated a large number of variants. Generally, chromosome I appeared to have a much higher rate of variation for its size, followed by chromosomes III, IV and II. Notably, 3 of the AUUH genes are on chromosome IV.

In order to observe mutations within the heme biosynthesis pathway, I then mapped the reads to the human coding sequence of each humanized pathway member. I repeated this mapping for genes which were not humanized, by using their yeast CDS as the reference. After mapping, consensus sequences for each gene in each sample were generated. For humanized genes, the consensus sequences of ancestral AUUH and the other 4 evolved strains were aligned to each other. For non-humanized genes, the BY consensus was included as well.

The alignments were largely identical to the reference sequence, with rare exceptions: The most striking difference was in UROS of strain 12D, where a 4 bp insertion after residue 247 causes a frameshift affecting the remaining

17 residues at the C-terminus of the protein (Figure 4.13). The frameshift also appears to have abrogated the normal STOP codon. Mutations at this residue[118], and further in the disrupted region[119], are known to be associated with human porphyrias. P248Q is associated with a porphyria phenotype in humans, where ALAS2 gain-of-function partially rescues the hypoactivity of UROS. Therefore, it is interesting and notable that this mutation should arise in our suppressor screen.

Another terminal mutation is observed for HMBS in strain 12C. From residues 352 onwards, there are several ambiguous nucleotides for the rest of the terminus (4.14). Though this could potentially indicate a significant mutation, in every case one of the possibilities for the ambiguous nucleotides is always the wild type. There do not appear to be prominent known mutations in this region; the most C-terminal known mutation I could locate is at residue 343.

HEM15 in 12C has two ambiguities near the N-terminus of the protein sequence (Figure 4.15). These are likely to be spurious. Firstly, only a few reads report an alternate base in those positions and the ambiguity appears to have been caused by coincidentally poor read quality and coverage in this spot. The N-terminus of HEM15 has a very significant role - as our earlier work demonstrates, the first 100 bp of the iron chelatase sequence contain critically important localization signals. Disrupting localization causes slow growth and harmful accumulation of porphyrins. However, in this case, even if the ambiguities were indeed mutations, one would be changing a leucine

to isoleucine, the other would be synonymous. Therefore, it appears that no significant mutation has occurred in the 12C strain at HEM15.

4.5 Mating array of humanized heme strains

In order to probe the epistatic landscape of the partially humanized heme biosynthesis pathway, and demonstrate the high-throughput gene drive cross concept, I decided to conduct a large mating array. Throughout my experiments, I had performed many CRISPR transformations in duplicate, for BY4741 as well as BY4742, thus I had humanized strains in a and α mating types. I collected 10 MATa strains (the "query strains") and 6 MATα strains(the "subject strains"). To these, BY4741-2 were added to serve as controls, for a final set of 11 query strains and 7 subject strains. These were distributed across subjects and queries of a 96-well plate to achieve an all-by-all cross and grown on appropriate selective media to select for successful mating on the basis of Lys and Met markers and CRISPR activity on the basis of Ura marker (conferred by the CRISPR plasmid).

Afterward, I compared the fitness of resulting diploids with a growth curve experiment in YPD. As a proxy for fitness, I calculated the area under the curve of OD measurements relative to the wild type, BY4741 mated to BY4742. Query strains mated to BY4742 and subject strains mated to BY4741 were then used as the baseline for establishing baseline growth rate for each strain. Expected relative growth rates were calculated as the product of baselines for either parent, as per the multiplicative model (Figure 4.17). There appeared to be substantial variations in observed growth rates. Some individual crosses stood out:

- HMBS to UROD and UROS grew very well, which is consistent with se-

quential transformation experiments where UU was modified into HUU.

- UROD to UROS was also a rapidly growing cross, similarly consistent with the construction of UU.
- Crosses with HMBS generally grew well. HMBS has also been quite easy to humanize in earlier experiments.
- Crosses to the evolved AUUHe seemed to grow slightly better than ancestral AUUH, which would be expected due to the higher fitness of AUUHe. Notably, FECH x AUUHe (but not FECH x AUUH) grew slightly better than expected, while ALAS1 x AUUH (but not ALAS1 x AUUHe) grew worse than expected. Both FECH and ALAS1 had proved recalcitrant to humanization in the context of the ancestral AUUH strain. However, given the superior growth of crosses to the evolved AUUHe, humanizations might prove more successful in this more optimal genomic context.
- AUU x HMBS grew much better than HUU x ALAD, which is also consistent with earlier observations.
- Crosses with subject strain PPOXk grew slightly better than expected, while PPOXdel crosses grew slightly worse. HEM14 is the yeast ortholog of PPOX, and it overlaps BUD25 on the opposite strand. BUD25 is non-essential. We generated two versions of the replacement template for PPOX, one which deletes part of BUD25 coding sequence (PPOXdel) and one which attempts to leave it intact. BUD25 is not known to interact with any of the heme biosynthesis pathway genes, but the diploid mutant has defects in bud-site selection[120] and might plausibly impact

the growth of diploid strains.

- Subject strain ALAD and query strain UROD generally showed poor growth rates, which is not consistent with the generally high level of fitness observed for these strains in earlier work. Therefore, the data for these two strains might have an unusually high level of error.

Figure 4.18 shows some summary visualizations comparing different groups of crosses. First, I sought confirmation of the multiplicative model which would predict that fitness of multiple humanizations follows the product of individual humanizations. In this experiment, the observed growth rates varied substantially from this prediction, so it seems likely that the effects of humanization are not independent. This supports the idea that inter-pathway interactions are a significant determinant of swappability. I also compared crosses with evolved AUUH and ancestral AUUH. Generally, the crosses with evolved AUUH fared much better. This is consistent with evolved AUUH growing better in previous experiments.

I also considered crosses with multiply humanized strains and singly humanized strains. In sequential experiments, there seemed to be a trend of gradually diminishing fitness as more humanizations accumulate. In this case, growth rates of crosses between singly humanized strains were slightly higher than with multiply humanized ones, but generally, there was not a clear tendency to be greater or less than the expected value under the multiplicative model. Lastly, comparing humanizations which I had earlier failed to obtain in combination, and those which were part of the AUUH quarter, there was

likewise little difference. The variation for successful genes was greater, however. This seems to recapitulate the process of building AUUH: At each step, several genes were transformed and only some were successfully added to the strains. Likely, these successful experiments coincided with combinations that have unusually good fitness. This explanation would predict, then, that positive epistatic relations contributed to the viability of AUUH; even though it grew quite slowly, other attempted quadruple humanizations could not be obtained at all.

4.6 Conclusion

In this chapter, I have presented my latest work on yeast humanization, in the context of humanizing heme. Being the most recent, this work is also in the earliest stage. Nevertheless, there appear to be some promising results:

- Construction of the quadruple-humanized heme strain was remarkably straightforward. Though the resulting strain was slow growing, it was amenable to manipulation and would be a suitable organism for large assays.
- Evolving the AUUH strain has been a very effective strategy. All 8 lineages have quickly recovered more than half of their growth defect. This is exciting for two reasons: It shows that humanization or other swapping experiments can be greatly facilitated by simple evolutionary approaches. The improvement in growth rate that I observed came after only 12 passages over as many days, with only background mutagenesis driving sequence diversity. Also, sequencing has revealed many more genetic variants arising in the evolved lineages, demonstrating that the suppressor screen is an effective approach to studying humanized pathways.
- Sequencing, in conjunction with evolution, has identified mutations associated with human disease. This underscores the clinical relevance of humanization studies.

The compatibility hypothesis, regarding whether interactions within a pathway or outside play a more significant role in humanization, remains elu-

sive. Sequencing data appears to imply that human genes specifically are being mutated, as no hint of mutation was observed in the non-humanized genes. Meanwhile, the mating crosses show very poor correspondence to the multiplicative model. More precise studies are needed to answer these questions – for example, introducing a mutagen into the evolution experiment could increase sequence diversity, and perhaps bring more heme suppressors above the detection threshold. Meanwhile, an evolved AUUH is an excellent candidate for introducing mutant human alleles and observing their sequence changes.

4.7 Materials and Methods

4.7.1 Strains and culture conditions

All strains were engineered from BY4741 and BY4742 strains of *S. cerevisiae*[121]. When selection was not needed, cells were grown in yeast peptone dextrose medium (BD, 242820). Selection was performed in SD-Ura, SD-Lys, SD-Met, CSM-Lys-Met, SD-Lys-Ura and SD-Met-Ura media (Sunrise Biosciences; 1703-500,1745-300, 1740-100, 1345-030, 1018-010, 1092-010). Media were supplemented with 2% agar for solid culture. Strains were suspended in 20% glycerol and frozen in -80 °C for long term storage.

4.7.2 CRISPR transformation

Sequential construction of the AUUH strain was done with sequential transformations of a CRISPR/Cas9 system as described in [21]. Briefly, a linear repair template DNA was constructed by PCR amplification and column purification (Zymo DNA Clean&Concentrator-25 kit, Zymo Research, D4005) of the human ortholog with homologies to yeast promoter and terminator sequences inserted via primers. A CRISPR plasmid targeting only the yeast coding sequence (CDS) was prepared with the Mo-Clo plasmid toolkit[23]. Cells were grown to OD=1 in liquid yeast peptone dextrose medium (YPD). Usually, overnight culture was sufficient, but for slow growing cells, sometimes 2-3 days of culture were required. The culture was made chemically competent using the Zymo EZ Yeast Transformation II kit (Zymo Research, T2001). The competent cells were transformed with 5 µg of the repair template and 500

ng of the plasmid, then plated on SD-Ura agar plates and cultured at 30 °C until colonies appeared. Reagents used were the same as those constructed and described in [37].

4.7.3 Plasmid curing

Two strategies were used for curing plasmids: When rapid curing of many strains was necessary, they were grown for 2 days on 5-FOA medium [105] and the resulting cells were free of the Ura auxotrophy (as confirmed by failure to grow on -Ura medium). For smaller sets of strains, the cells were streaked out on YPD agar plates and incubated until colonies appeared. Then 12 numbered colonies were patched onto a fresh YPD agar plate as well as -Ura plate and cultured in parallel. Colonies which failed to grow on -Ura after 6 days were considered cured of the plasmid.

4.7.4 Colony PCR screen

Up to 12 colonies at a time were picked from transformation plates and probed for presence of the human gene by PCR [21]. The PCR confirmation primers were designed so that one primer anneals to the yeast promoter which was not covered by the repair template homology, while the other binds to the human coding sequence and not the yeast one. Expected product sizes ranged between 200-500 bp. Primer sequences are the same as those in [37]. PCR was analyzed by gel standard agarose electrophoresis.

4.7.5 Agar plate suppressor screen of AUUH

A clonal stock of AUUH colony was grown overnight on YPD agar. The cells were collected and suspended in water for OD measurement. The cells were diluted to ∼10 million CFU/ml based on OD (0.68 OD is assumed to correspond to 10 million CFU/ml). From this master dilution, serial dilutions of 1:10 were made in 1 ml water each. 300 µl of cells from several dilutions was plated on 150 mm YPD agar plates for 30, 300, 3k, 30k expected colonies, to account for measurement and dilution error and cultured for 3 days. The plate with the largest number of separated colonies was selected to collect suppressors. 12 large and 12 small colonies were picked with sterile toothpicks and confirmed for AUUH humanization.

4.7.6 Evolution of AUUH

1 µl of a clonal population of AUUH from glycerol stock was inoculated into eight tubes with 5 ml liquid YPD media. Cultures were designated A-H to indicate lineage and numbered 1 for day 1. After exactly 24 hours of culture at 30 °C with orbital shaking at 200 rpm, each tube was removed from culture. Cells which had precipitated to the bottom were resuspended by pipetting up and down inside the tubes. Then, 1 µl of each tube was inoculated into a fresh 5 ml YPD culture to create cultures A2-H2. The remainder of A1-H1 was frozen in glycerol. This process was repeated for 12 days resulting in A12-H12. All cultures were saved in glycerol.

4.7.7 Growth curves

Yeast strains were pre-cultured in liquid YPD or -Ura medium overnight. 1 µl of each culture was inoculated into 200 µl liquid culture on 96 well format transparent plate. A Synergy H1 spectrophotometer (BioTek) was used to incubate the plates, with shaking and optical density measurement every 15 minutes for up to 72 hours. Data was analyzed in triplicate, except in the case of the pairwise mating array.

4.7.8 Genome sequencing of AUUH

For genome sequencing, samples were grown overnight in YPD and DNA was isolated with the Zymo YeaStar Genomic DNA Kit (Zymo Research, D2002). 3 µg of each sample was submitted to the CSSB sequencing core in 100 µl for 250 bp paired-end Illumina reads. Data were analyzed with Geneious v9[96] software. Reads were paired and mapped to the S288C reference genome from SGD[44] with the Geneious mapper. Mapping statistics were exported and analyzed in R with `dplyr` and `ggplot2` packages. Multiple sequence alignments were performed with the MUSCLE software[122, 123].

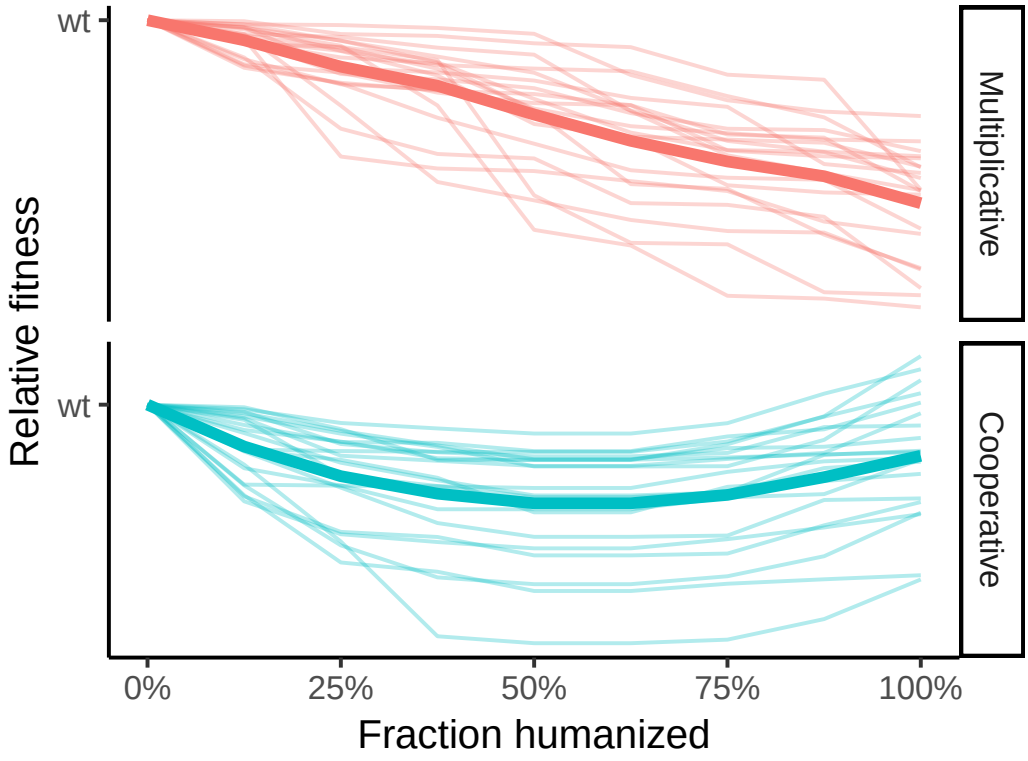

Figure 4.1: Illustrative simulation of the multiplicative and cooperative models for high order humanization. Each thin line shows fitness over the hypothetical course of a whole-pathway humanization project for a module with 8 genes. Averages of 20 runs are shown as thick lines. **(Top)** Under the multiplicative model, fitness f_i of humanizing gene i is assumed to be independent of other humanizations in this module, thus the final fitness is given by the product $\prod f_i$ calculated over every i humanized in that strain. In this simulation, $f_i = 1 - x$ where x is an exponentially distributed random variable with mean 0.05 ($\lambda = 20$). **(Bottom)** Under the cooperative model, f_i is strongly dependent on the whether other genes in the module have been humanized. In reality, a complex epistatic network of $n \times n$ genes would determine the fitness of every subsequent humanization. For simplicity, in this simulation $f_i = 1 - x \cdot (0.5 - r)$ where x is an exponentially distributed random variable with mean 0.1 ($\lambda = 10$) and r is the fraction of genes in the pathway that are already humanized.

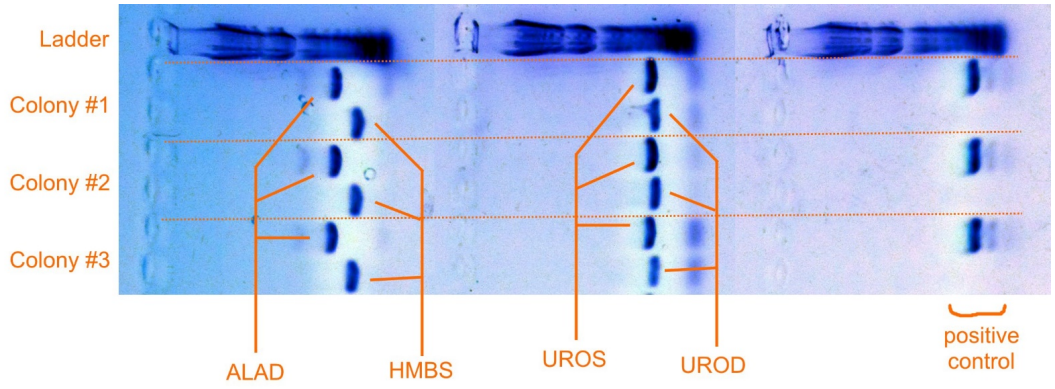

Figure 4.2: Colony PCR confirmation of quadruple humanized AUUH strain. This gel shows agarose electrophoresis and SYBR Safe staining colony PCR for 3 different colonies from a CRISPR humanization experiment. From top to bottom, the first row is a GeneRuler 1 kb DNA Ladder. The next three pairs of lanes correspond to colonies 1, 2, 3 respectively. Bands corresponding to PCR products spanning the junction of the humanized hem2::ALAD, hem3::HMBS, hem4::UROS and hem12::UROD and their yeast promoter regions are shown. All 4 bands are absent when the humanization is not present as one primer in each pair binds only the human and not the yeast sequence (not shown). The rightmost lane is a positive control (the yeast ERG13 locus). All bands are at their expected size range.

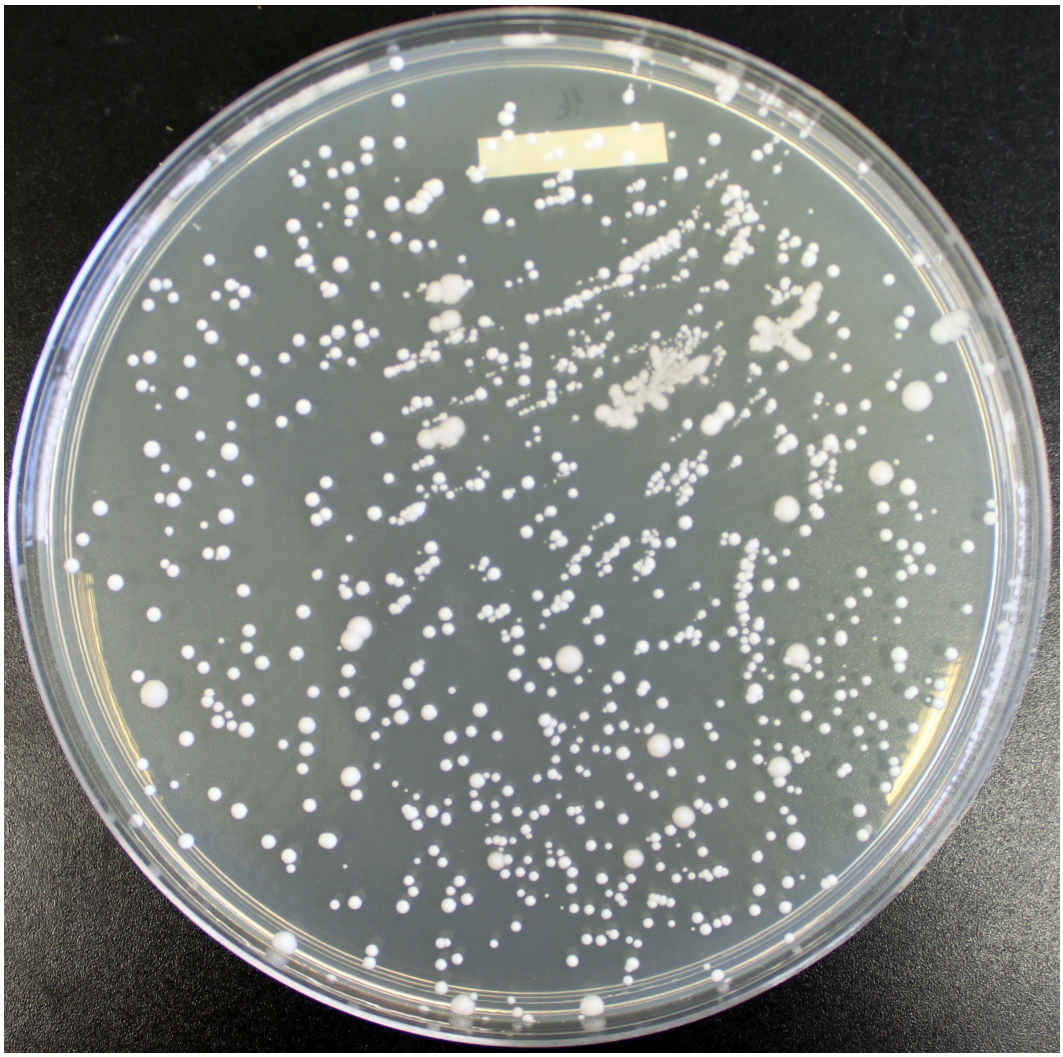

Figure 4.3: AUUH suppressor screen feasibility study. AUUH cells from a clonal population were cultured on YPD agar for 3 days, giving rise to this pattern of colonies. Large and small colonies can be seen.

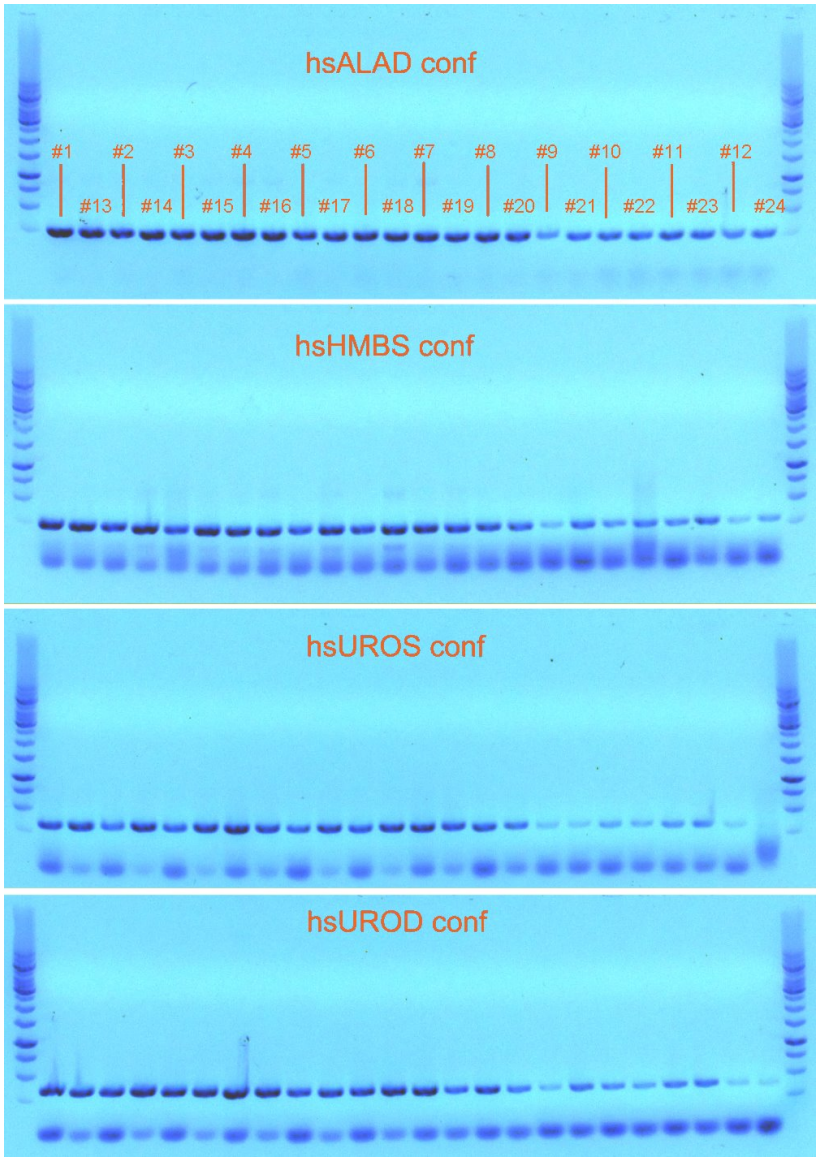

Figure 4.4: Colony PCR confirmation of AUUH suppressor colonies. Lanes, from left to right, are large (1-12) and small (13-24) colonies from the AUUH suppressor screen on solid medium. Each row shows results of probing for a different gene. All 4 humanizations are shown to be present in all 24 colonies.

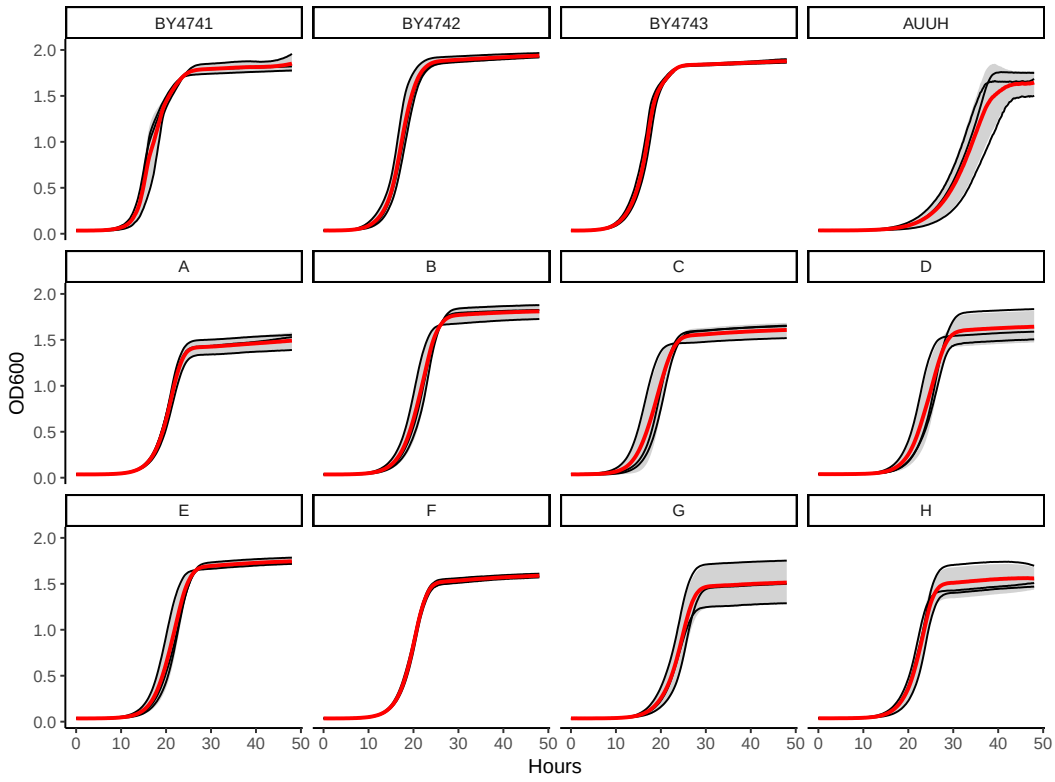

Figure 4.5: Growth curves of individual evolved AUUH lineages. Each plot shows measurements of three parallel cultures. Black lines show individual OD measurements, red line is the mean OD and grey bands show standard deviation ($\pm 1\sigma$). "AUUH" denotes the ancestral AUUH strain which was not passaged, while plots labeled A-H are 12th day samples of each evolved lineage.

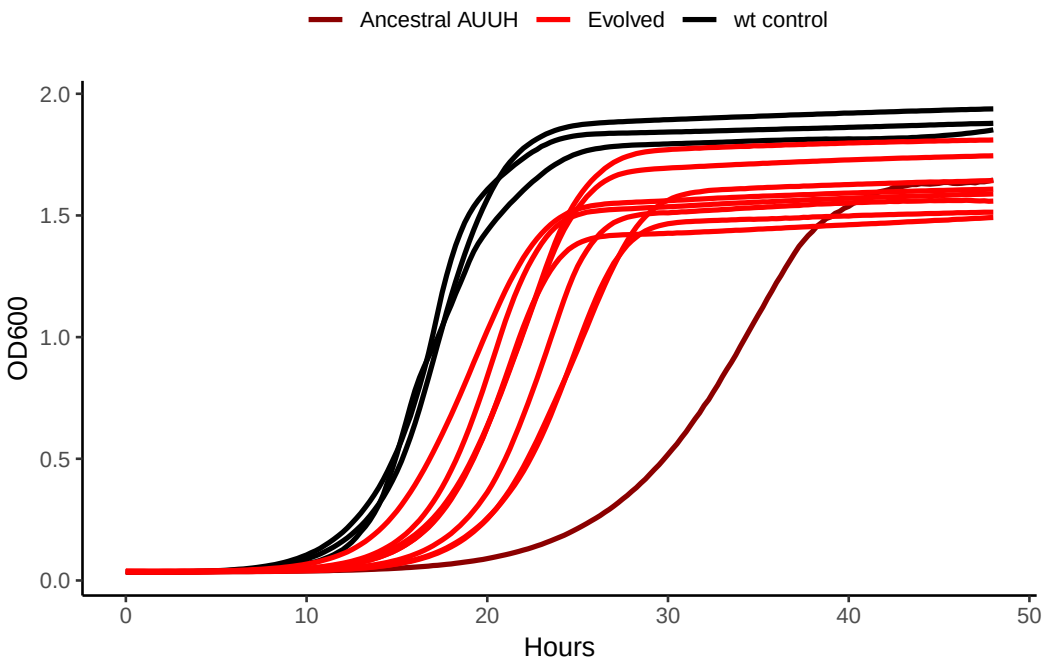

Figure 4.6: Comparison of mean growth curves for evolved AUUH lineages. Red lines show 8 evolved lineages, dark red shows the ancestral AUUH and black lines show 3 wild type cultures (BY4741, BY4742 and BY4743).

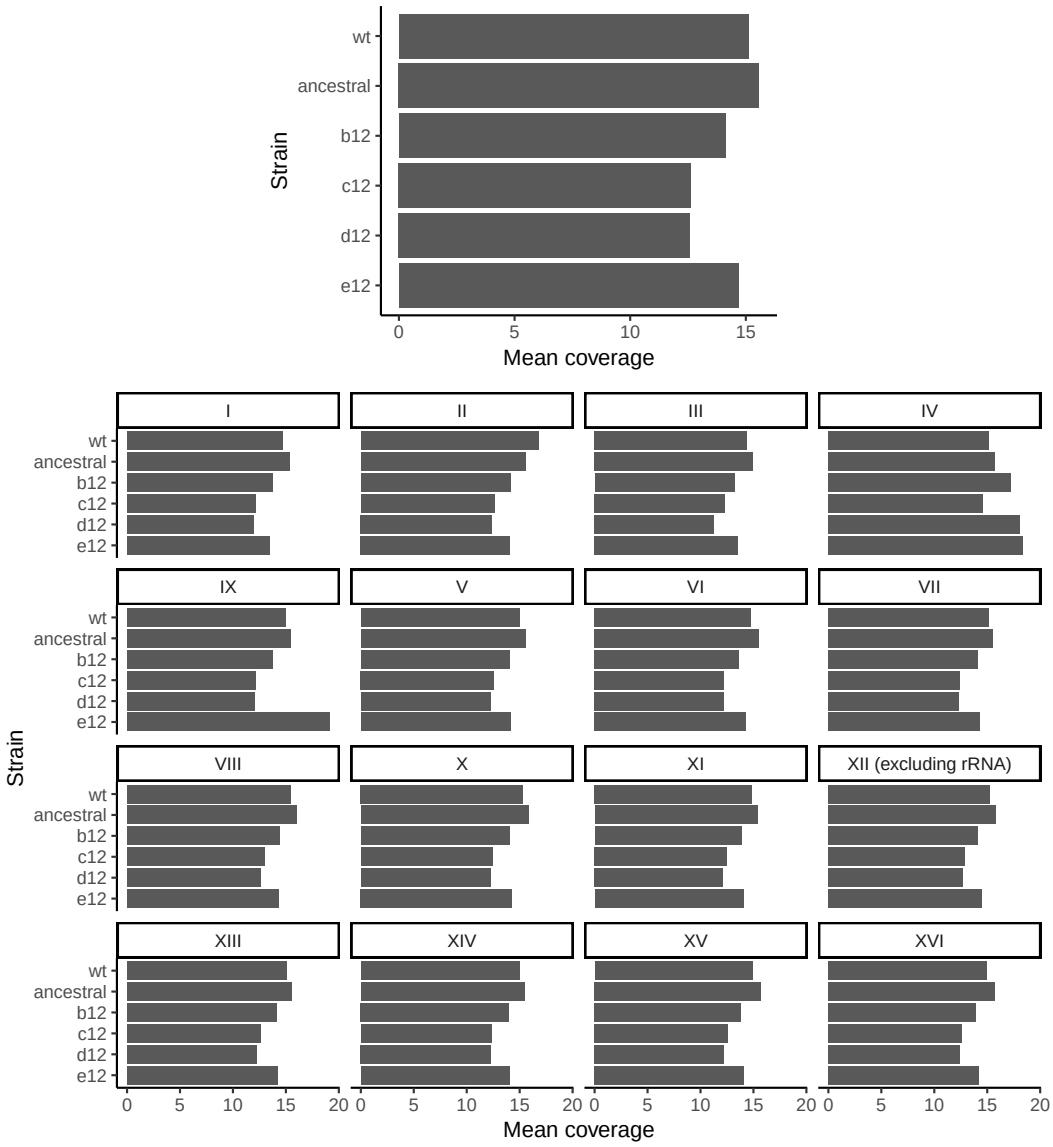

Figure 4.7: Mean coverage of whole genome sequencing reads. Mean coverage is shown across all chromosomes **(top)**, and by individual chromosomes **(bottom)**. For chromosome XII, reads mapping to the highly repetitive ribosomal DNA sequence (Figure 4.8) were excluded from coverage calculations.

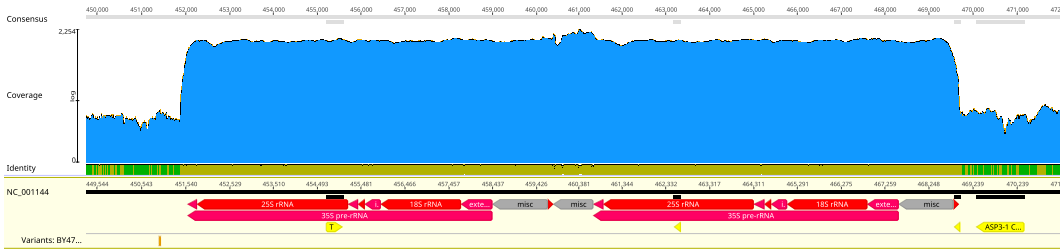

Figure 4.8: Very high apparent coverage of ribosomal DNA on chromosome XII. Many repeats of ribosomal DNA are represented as a single locus in the reference sequence for chromosome XII. Shown here is coverage resulting from mapping of BY4741 reads. Coverage is on log-scale, and is about 1200x within the rDNA.

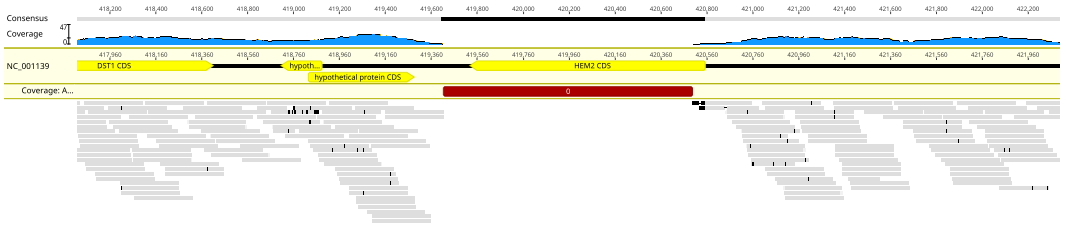

Figure 4.9: Coverage gap around genomic HEM2 in AUUH. HEM2 was replaced by ALAD.

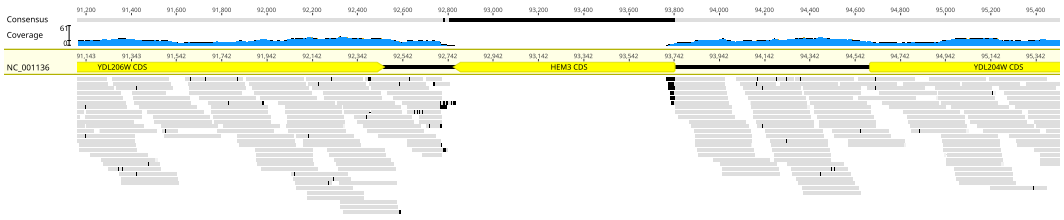

Figure 4.10: Coverage gap around genomic HEM3 in AUUH. HEM3 was replaced by HMBS.

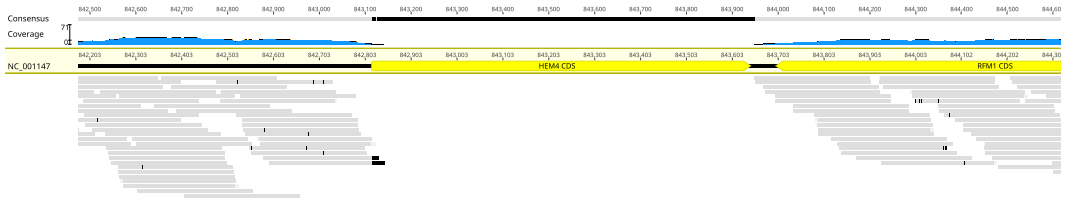

Figure 4.11: Coverage gap around genomic HEM4 in AUUH. HEM4 was replaced by UROS.

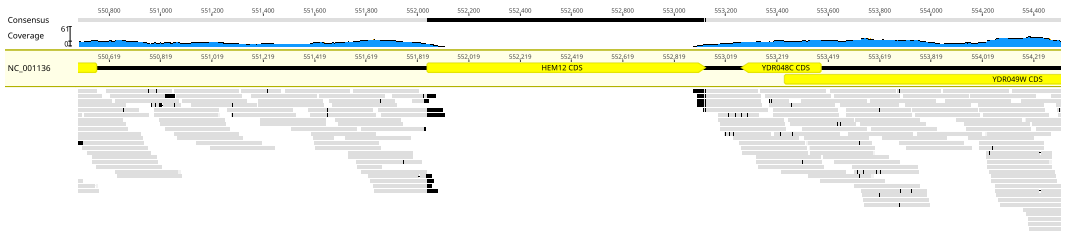

Figure 4.12: Coverage gap around genomic HEM12 in AUUH. HEM12 was replaced by UROD.

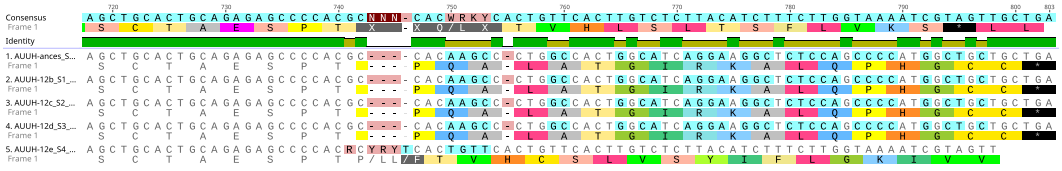

Figure 4.13: Multiple alignment of consensus UROS sequences from AUUH strains. A 4 bp insertion leads to a frameshift covering roughly the last dozen residues of the protein. Two known disease mutations occur in this region, one of them almost exactly at the location of the insert.

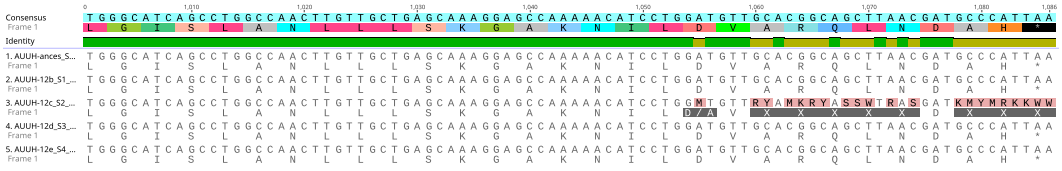

Figure 4.14: Multiple alignment of consensus HMBS sequences from AUUH strains. Ambiguous nucleotides in the C-terminus imply a possible mutation. The individual reads have a bimodal character and seem to follow two distinct sequence patterns. The alternate pattern would cause a frameshift.

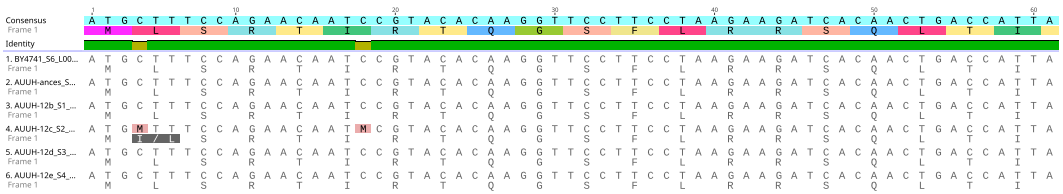

Figure 4.15: Multiple alignment of consensus HEM15 sequences from AUUH strains. Two ambiguous nucleotides may potentially be causing residue changes. The first, M indicating canonical C or mutant A, would lead to an L2I mutation. The second, also an M with canonical C, would be a synonymous substitution. However, there are not many actual reads exhibiting the alternate sequence.

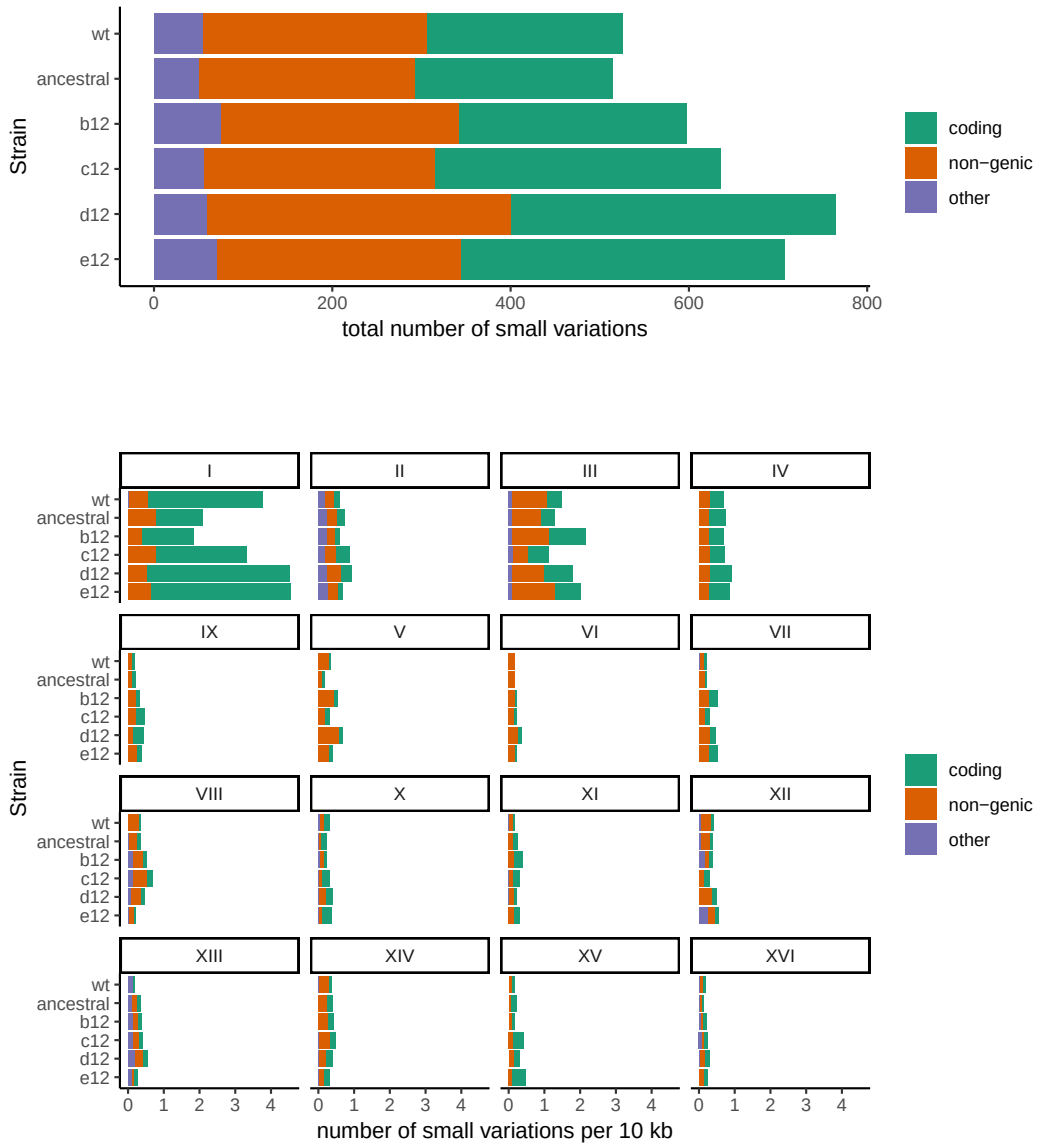

Figure 4.16: Total counts of small variations vs. reference yeast genome. **(Top)** Absolute counts across entire genome. **(Bottom)** Counts in individual chromosomes, normalized to chromosome size.

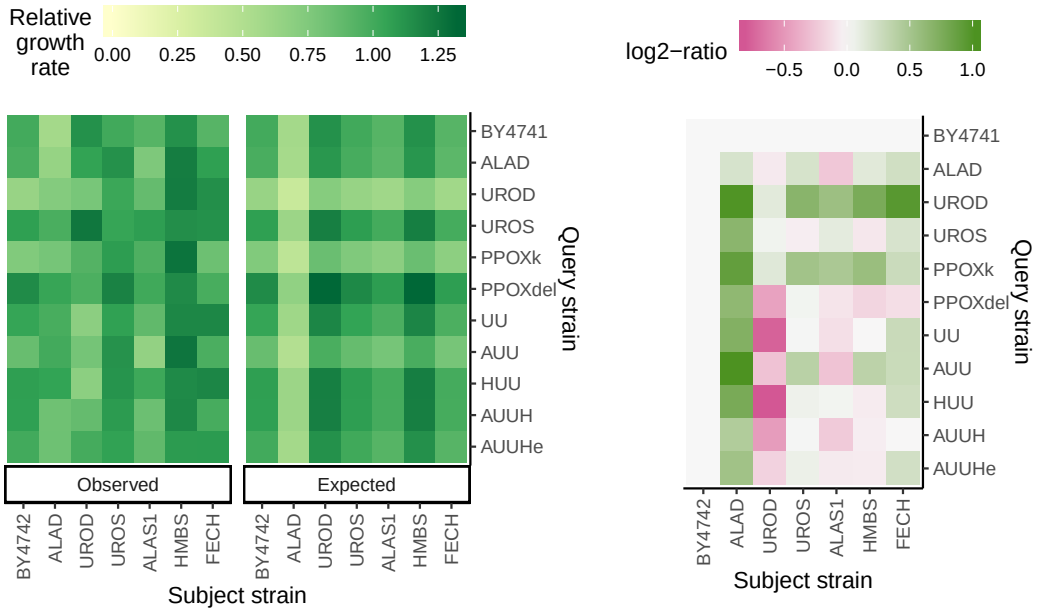

Figure 4.17: Observed and expected growth rates of various crosses between humanized strains. **(Left)** Heatmaps showing area under curve relative to the BY4741 x BY4742 cross. **(Right)** Log-ratio comparison of expected to observed. Calculated as $log2\frac{\text{AUC}_{\text{obs}}}{\text{AUC}_{\text{exp}}}$.

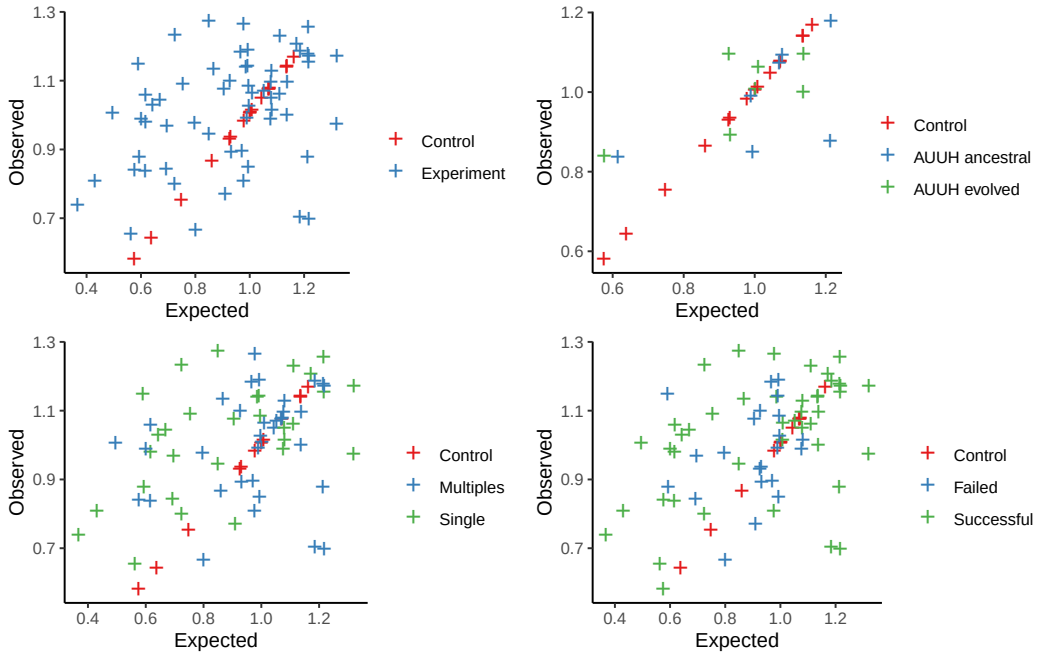

Figure 4.18: Scatterplots summarising relative growth rates for selected groupings of crosses. "Expected" refers to the growth rate predicted as the product of crosses to BY4741-2 (also shown in red). **(Top left)** Overall, there was substantial variation in growth rates, with more crosses growing better than expected. **(Top right)** Crosses with evolved AUUH were more likely to have better than expected growth rate, and grew higher overall. **Bottom left** Crosses with singly humanized strains were more likely to grow better than expected as opposed to crosses to multiply humanized strains, but generally, the difference was slight. **Bottom right** There was not much difference between growth rates of crosses with humanizations that had previously been obtained in combination (ALAD, UROD, UROS, HMBS) and those which were not (FECH, PPOX, ALAS1).

Chapter 5

A Highly Parallel Strategy for Storage of Digital Information in Living Cells

DNA has a long history as a medium of storing information, but humanity has only began to explore this potential recently. Initial investigations into the possibility of writing digital data into DNA sequences were confined to theory, and it is only in the recent past that the advent of high-throughput sequencing and synthesis technology made arbitrary reading and writing of DNA possible. Several prominent researchers have produced systems for doing this[125, 124, 126, 127], but they appeared unanimous in a key choice as to how the information is to be partitioned.

Though modern electronic memory can fall short in very long term survivability, it has succeeded in providing vast storage space. Computer files can often be thousands, millions or even billions of symbols long as the complexity of software grows. Conversely, DNA synthesis technologies are limited. It is not practical to synthesize pieces of DNA much longer than a few hundred basepairs, especially on the vast scale needed for information storage. Thus

This chapter was previously published in Akhmetov A, Ellington AD, Marcotte EM (2018) *BMC Biotechnol.*, 18(1):64. EMM helped conduct the research. EMM and ADE also helped me conceive the main idea and write the manuscript.

any system for encoding information in DNA must first overcome the hurdle of breaking that information down into tiny pieces, and keeping track of the pieces. Upon my first introduction into this area of research, I noticed that all existing approaches seemed to favor a strategy of embedding individual addresses into each piece. This method has some advantages (conceptually straightforward, random-access) but also disadvantages (risk of address corruption, complexity of implementation). My work was motivated by the observation that it is possible to use a wholly orthogonal approach.

I developed a proof of concept codec which uses a customizable algorithm to generate a codebook for a perfect code. The user is able to directly specify a desired block size and error tolerance capacity. I further developed a method of partitioning the resulting information without addresses, but rather with overlaps between successive packets. As the problem of assembling many small pieces into one long sequence has been solved extensively by traditional de novo sequencing, my approach is very compatible with off-the-shelf sequencing software rather than requiring a specialized assembly solution. The overlaps act as an additional protection against error. I tested the performance of my codec by simulating a round-trip synthesis and sequencing of digital information in DNA, with results demonstrating the feasibility and practicability of this method. Besides performing competitively with regards to other studies, my approach has an additional advantage that is very significant from a very long term archival perspective: It requires the storage of minimal details regarding its operation for successful recovery to be possible. It also relies on

technologies with inherently enduring interest to humanity, such as sequencing of large genomes. Because of this, information encoded with my method is at much lower risk of getting lost due to technologies needed to read it becoming forgotten or obsolete (a common problem of many storage systems).

5.1 Abstract

5.1.1 Background

Encoding arbitrary digital information in DNA has attracted attention as a potential avenue for large scale and long term data storage. However, in order to enable DNA data storage technologies there needs to be improvements in data storage fidelity (tolerance to mutation), the facility of writing and reading the data (biases and systematic error arising from synthesis and sequencing), and overall scalability.

5.1.2 Results

To this end, we have developed and implemented an encoding scheme that is suitable for detecting and correcting errors that may arise during storage, writing, and reading, such as those arising from nucleotide substitutions, insertions, and deletions. We propose a scheme for parallelized long term storage of encoded sequences that relies on overlaps rather than the address blocks found in previously published work. Using computer simulations, we illustrate the encoding, sequencing, decoding, and recovery of encoded information, ultimately demonstrating the possibility of a successful round-trip read/write. These demonstrations show that in theory a precise control over error tolerance is possible. Even after simulated degradation of DNA, recovery of original data is possible owing to the error correction capabilities built into the encoding strategy. A secondary advantage of our method is that the statistical characteristics (such as repetitiveness and GC-composition) of en-

coded sequences can also be tailored without sacrificing the overall ability to store large amounts of data. Finally, the combination of the overlap-based partitioning of data with the LZMA compression that is integral to encoding means that the entire sequence must be present for successful decoding. This feature enables inordinately strong encryptions. As a potential application, an encrypted pathogen genome could be distributed and carried by cells without danger of being expressed, and could not even be read out in the absence of the entire DNA consortium.

5.1.3 Conclusions

We have developed a method for DNA encoding, using a significantly different fundamental approach from existing work, which often performs better than alternatives and allows for a great deal of freedom and flexibility of application.

5.2 Background

Recently, the prospect of encoding information in free nucleic acids has attracted much interest from both academic research communities[124, 128, 125, 126] as well as the technology sector[129]. DNA offers unique potential for storage of information, in that large amounts of information can be be written (synthesis) and read (sequencing) at moderate, and rapidly decreasing, cost. Ultimately, one DNA base-pair (bp) stores 2 bits of information [124], a much more dense information storage medium than any electronic device of comparable capacity. Moreover, the long term storage of information in DNA is potentially very feasible, given its extremely long half-life[130] unlike digital media which is prone to degrading with timescales on the order of decades. As an example, it has been possible to reconstruct a mammoth genome from remains found in the tundra[131], it is unlikely we would recover electronic information stored in the same way. This feat was possible in part because an exquisite molecular mechanism (base-pairing and replication) exists for making many copies with very high fidelity.

Herein we propose a novel scheme for encoding information in DNA and distributing this information across multiple cells, and present computer simulations demonstrating the feasibility of our approach. We discuss three main steps of this process (Fig.5.1): (i) A coding scheme for converting digital information into DNA and vice versa. Our scheme has a built-in, highly general error detection and correction capacity that can be tailored to pre-chosen balances of redundancy versus error tolerance; (ii) A strategy for parceling long

strings of information into smaller pieces that allows for their later re-assembly. This strategy is completely compatible with current chemical DNA synthesis methods that yield at most only short ($\tilde{2}$00 bp) oligonucleotides (oligos); and (iii) The use of error tolerance to defeat corruption of information arising from synthesis errors, sequencing errors, mutations and packet loss. Taken together, these innovations allow for very long term, error-resistant storage, potentially for thousands of years.

We also analyze the performance of our method for key trade-offs that will be inherent in any strategy of encoding digital information in DNA.

5.3 Results

5.3.1 Successful generation of codebook

We executed our implementation of the codebook generation algorithm to generate a set of codewords 4 bp long, and with minimum Levenshtein distance[132] of 3 (the full set of parameters is given in Table 5.1: Parameters used for codebook generation). The latter parameter was set so as to allow recovery from up to one mutation (including substitutions, insertions and deletions) per block of encoded DNA. Due to a codeword length of 4, the block length is likewise 4 bp, therefore the expected theoretical upper bound on mutation rate for error recovery is 0.25 bp^{-1}, and the upper bound for error detection is 0.5 bp^{-1}.

All of the resulting code words showed a good diversity of base pair composition, lack of repetition, sufficient sequence distance between themselves, and overall conformed to expectations. The list of codewords is given in Table 5.2 along with the numeric value assigned to each codeword for the work described here.

The theoretical upper bound on the information that can be stored using the four nucleotides of DNA is 2 bits bp^{-1}. Given our codebook, each sequence of 4 bp can only have one of 8 values, therefore the information content under our encoding scheme is only 3 bits per 4 bp, with a density of 0.75 bits/bp. Thus, the expected theoretical rate of our encoding approach per se can be calculated as $\frac{0.75}{0.375} = 0.375$.

5.3.2 Encoding of digital data into DNA

We implemented our encoding algorithm and used it to encode five separate sets of input data (Table 5.3). Table 5.4 describes the performance of our encoding algorithm on these test data sets. Each input file was wrapped in a tar archive prior to encoding, and rates were calculated by dividing the size of the tar file prior to converting (found by multiplying the size in bytes by 8) by the information content of the resulting DNA (found by multiplying the number of base pairs by 2).

5.3.2.1 Cat image

The realistic test input Cat.jpg (Fig. 5.2) was encoded and had an empirical rate of 0.459, thus the redundancy and overhead introduced by our encoding algorithm inflated the size of the data by 1.179 times. Notably, this is a rate higher than the predicted 0.375; this is due to the Lempel-Ziv-Markov chain algorithm (LZMA) compression step built into our method reducing the size of the input (compression with LZMA yields only 10,028 bytes, which would result in a rate of 0.375). After decoding, we were able to recover the original image exactly.

The resulting DNA string appeared to be free of any major self-similarity or repetition. We visualized the self-similarity using a dot plot to show the extent to which the sequence matches itself (Fig. 5.3). The most notable case of repetition was near the beginning and ends of the sequence (Fig. 5.4). Closer inspection of the binary data after compression

but prior to transformation into DNA revealed that there is often a short string of zeros in LZMA-compressed data, containing header/tail information used by the compression logic to identify the properties of a compressed data stream. Indeed, the most commonly repeated sequence at these regions was ACCG, which maps to 0 in our codebook.

5.3.2.2 Random data, centromere and flat file

Random data is often regarded as particularly difficult to compress, due to a lack of statistical tendencies in it which can be exploited by compression strategies. Indeed, upon compressing our 10 kb random file with 7zip, an open source implementation of the LZMA algorithm, we obtained a compressed file 10.2 kb in size. When encoded into DNA with our algorithm, the resulting file was only 1.278 times bigger than the input tar file.

The centromere was included as an example of highly repetitive information that has statistical properties representative of known repetitive DNA that is considered challenging to sequence and synthesize. As seen by the high rate, the compression step of our encoding process vastly reduced the size needed to store this data. The resulting DNA string had very similar composition and structure to the other test data, and should be no more difficult to synthesize or read with our approach than any other input.

Lastly, the flat file was dramatically reduced in size after compression, as shown by the extremely high rate. Due to the very short output sequence, the repetitive head/tail regions are clearly visible in the dot plot (Fig. 5.5), as

the data content of the end result is relatively small.

5.3.2.3 Base pair composition

We noted that the nucleotide composition was very close to an even split of 25% for each base in most of our encoded DNA (Fig. 5.6). The most conspicuous exception was the flat file, which showed larger skews due to the characteristic head and tail patterns, which do not vary in length with the amount of data encoded, thus having a disproportionate effect.

We have also investigated the local composition to detect any small stretches of base composition skew (Fig. 5.7). As expected, we did not observe any regions of pronounced skew, and the distribution of nucleotides seemed to be uniform throughout the encoded sequences. Notably the skew caused by characteristic head and tail patterns is most noticeable in the flat file, where several peaks of C content can be seen at the beginning and end. This would be the expected result, since the common code word in these areas is ACCG which has a larger proportion of Cs. After quantifying the total skew by adding up local deviations from expected even distribution, we observed mostly small amounts of error, with the exception of the flat file which had larger composition imbalances, particularly in C and T (Fig. 5.8).

In order to verify that our encoding algorithm performs as expected for larger files, we have also performed a round-trip read-write of a larger image, Cat-big.jpg (Tables 5.3 and 5.4). The algorithm performed very well for this larger dataset and we were able to recover the original input exactly. The

overall error in composition was even less than smaller datasets (Figs. 5.6, 5.7 and 5.8), due to slightly better performance of the LZMA compression for larger input data.

5.3.2.4 Open reading frames within the encoded DNA

Encoded DNA produced by our algorithm is expected to resemble random sequence in many respects. It is possible that large stretches of such sequence would coincidentally contain start and stop codons forming open reading frames. Our scheme anticipates that the information-bearing DNA would ultimately be stored in living cells, thus it is of interest whether and to what extent open reading frames (ORFs) appear in the encoded sequence. We analyzed the occurrence of such random ORFs, and a visualization is presented in Fig. 5.9.

The distribution of ORF lengths follows a pattern that would be expected a priori from random sequence. The shortest ORFs are most frequent (our ORF detection software looks for ORFs containing at least one codon besides start and stop). ORFs longer than about 30 residues or 100 bp are very rare, and none are longer than 83 residues. The start and stop codons are evenly distributed throughout the sequence, but slightly under represented near the ends of the DNA – possibly owing to the composition skew introduced by the characteristic head and tail regions resulting from our encoding approach. There does not appear to be any particular bias towards a single reading frame, nor towards a particular direction.

5.3.3 Decoding and error correction

We confirmed correct operation of our encoding approach by attempting to decode the DNA resulting from the encoding step. In every case we were able to recover the original tar archive, which when extracted produced the relevant file that was identical to the one encoded.

Having verified successful round-trip encoding-decoding of information we then sought to measure the performance of the error correction function. We simulated substitution mutations accumulating at a slow but steady rate over many generations, by repeatedly mutating the Cat.jpg so as to correspond to total numbers of mutations per block ranging from 0 to 2. As an indicator of data integrity, we compared sequence identity between the simulated mutant sequence and the original, before and after applying the error correction (Fig. 5.10). For a smaller number of mutations, the error correction is able to restore nearly the entire original sequence. As the density of mutations increased, eventually the error correction did become overwhelmed, but the data shows a very clear mutation buffering effect contributed by it.

As calculated in earlier sections, our error correcting code operates on a block-wise basis. Each individual block (in our case 4 bp long) can be recovered so long as no more than one mutation occurs within it. With larger numbers of mutations distributed randomly throughout the sequence, it becomes more probable that at least two mutations will fall very close to each other and coincide on the same block. Such an improbable event would potentially result in the loss of that block of information. We have observed that minor pertur-

bations to the encoded DNA do not significantly corrupt the stored data, and do not preclude its recovery. Furthermore, our strategy anticipates that the final encoded DNA string will be broken up into small, overlapping pieces in practice; therefore the mutations would be further removed at the assembly stage via a consensus mechanism.

5.3.4 Parallelized storage

We investigated the practical aspects of synthesis, storage and reading of digital data as DNA using our approach by simulating the round trip process. Basing our model on the assumption of DNA synthesis capability which allows the production of a pool of oligonucleotides 200 bp each (readily achievable with current technology), we generated a series of 200 bp "packets" of information, which tile the encoded DNA with a pre-defined amount of overlap between each two successive packets (shown in Fig. 5.11 for a 175 bp overlap). These packets can be synthesized as a mixed pool, stored, and sequenced using standard high-throughput sequencing technology, which would allow assembly of the full-length DNA sequence using only the overlaps, without necessitating the complicating addition of addresses and addressing schemes. In order to ensure even coverage at the termini, we generated successively shorter fragments at these areas.

5.3.4.1 Simulated sequencing and assembly

After generating packets with varying overlaps (75, 100, 125, 150, 175 bp corresponding to mean sampling densities ranging from 1.6 to 8) we simulated the outcome of sequencing this pool of oligos using the ART sequencing simulator (details in Table 5.1) at varying read depths (1x, 2x, 5x, 10x, 50x). Afterwards we attempted de novo assembly of the resulting FASTQ files, compared the longest resulting contig to the original sequence, and considered two key measures of sequence similarity: How much of the original sequence was present in the resulting contig (recall), and how much of the resulting contig matched the original sequence (precision).

Our measures of precision and recall correlated very closely with each other: With read depths of 5x or more, assembly could easily succeed even when overlaps between successive packets are small. Below this threshold, only comparatively dense tilings of packets could be assembled: With 8x sampling density, even 1x read density was sufficient to recover the original sequence. However, lower sampling densities performed very poorly with low read depth, and contigs assembled would be a fraction of the size of the original sequence, as well as having little similarity to it (indicating spurious assembly becoming the dominant process). A visualization of these findings is provided in Fig. 5.12. Based on these results, we decided to use a sample depth of 8x (175 bp overlap) and a read depth of 10x for subsequent work.

Notably, we simulated sequencing of a complex pool of short oligos. Simulated reads obtained from this pool were assembled naively; we did not

attempt to recover individual packets, and then assemble packets. Rather, due to the overlaps between successive packets, our assembler was able to seamlessly combine these reads into a single contig without additional intervention. We consider this to be an advantage of our conservative overlap-based segmentation of data.

5.3.4.2 Library construction and long-term packet-wise retention

Because the intended use of our strategy involves cloning the pool of synthesized DNA oligos, there is a risk that not every sequence in the pool will be represented. If too many packets "drop out", our assembly method may suffer catastrophic failure (a full contig will be impossible to build if a critical number of adjacent packets are missing entirely at one or more locations). However, with higher sampling densities, it is possible that even though one packet is lost during cloning, the two packets adjacent to it will bridge the gap and nevertheless rescue successful assembly. In order to investigate the validity of this tradeoff in the practical context of our work, we performed Monte Carlo simulations of library coverage under the assumption of a defined number of clones being harvested (set to a constant multiple of the sequence diversity).

In all, we conducted 20 simulations each from pools of 100 and 500 unique sequences and expected mean library sample rates ranging from 1 to 30 clones harvested per unique oligo sequence. We then looked at what fraction of the initial set was represented in the draw (Fig. 5.13). As expected, if the

total number of clones harvested is equal to the number unique sequences, some sequences appear repeatedly and many are not captured at all. In our case, about a third of the pool would be lost under this cloning regimen. Collecting a very large number of clones virtually guaranteed that no sequence would be lost. Interestingly, the threshold of full coverage was between 10 and 5 clones per unique oligo (cpo): The vast majority of the sequences could be recovered with 5 cpo, but in most simulated runs there would be a few packets missing. On the other hand, if 10 clones per oligo are considered then every single packet was recovered in all 20 simulation runs.

Interestingly, the complexity of the pool per se does not appear to have an effect on these boundary conditions. The most pronounced difference between the 100 oligo and 500 oligo case was that the variation was greater with a less complex pool, while the more complex pool behaved more predictably in individual runs.

We concluded, based on our observations, that harvesting 10 clones per oligo is sufficient to have a >95% confidence that all original packets will be well represented.

5.3.4.3 Simulated recovery of information with packet loss

Having observed that harvesting roughly 10 clones for each unique oligo should almost guarantee full library coverage, we attempted to test this inference with in silico experiments. We conducted a series of simulated experiments in which the initial pool of oligos was randomly sampled with replace-

ment (since the number of molecules in each class is typically much higher than the number of distinct sequences in synthesized oligo pools, we regarded the effect of replacement as negligible). These random subsamples were then subjected to simulated sequencing with ART at 10x read depth, and then de novo contig assembly of the resulting reads was attempted to determine whether recovery of the original sequence is possible even with lost packets. This was repeated 20 times to account for the influence of chance events on recovery.

As shown in Fig. 5.14, results were in line with the expectations arising from the library-coverage experiments. We saw that with a single clone harvested per oligo, in all 20 cases the resulting contigs do not match the original sequence. Moreover, we have seen that due to very poor coverage of the original pool, under this regimen the assembly suffers catastrophic collapse: The resulting contigs do not exceed roughly a third of the original sequence by length, and their sequence often diverges from the reference. Predictably, applying the error correction did not ameliorate this situation.

With the previously established "safe" regimen of 10 clones per oligo, the vast majority of the 20 experiments yielded contigs that were exactly identical to the original sequence, demonstrating successful round-trip read-write of digital data as DNA. In two cases, the assembled contig did not match the original exactly, but the only difference in either case was a single missing terminal nucleotide. Because our error correction method is based on Levenshtein distances, which take into account not only substitution but also deletions, it

was possible to correct these problems and restore the original sequence exactly. Thus, considering error correction as well, with a regimen of 10 clones per oligo we could reconstruct the original sequence exactly in every single one of our 20 experiments.

Lastly, of interest was the borderline case of 3 clones per oligo. As expected, this regimen was occasionally able to produce exactly matching sequence, but often the contigs would differ slightly due to missing one terminal nucleotide. Many such errors were readily amendable with the error correction, such that attempting to correct errors produced a dramatic improvement in how many assembled contigs matched the original sequence exactly.

5.4 Discussion

DNA has great potential as a medium of information storage. Indeed, it has been used for this purpose by all living organisms for millions of years. Molecules of DNA are much smaller than digital devices, can be easily copied using both natural and artificial systems and they can be stably maintained for a very long time: The half-life of DNA in solution, depending on pH, temperature, and length, can range from a few years to hundreds of thousands of years[133]. In nature, DNA in fossilized bone has been shown to have a half-life as long as 500 years, implying recovery should be possible after many thousands of years given sufficient initial copies of the DNA[130], and useful sequence has been recovered from 450 to 800 kyr old samples encased in ice[134]. In contrast, digital media degrades quickly over decades[135, 137, 136, 138].

Besides durability of the medium itself, DNA enjoys a unique advantage in that the characteristics of DNA are fixed over time. In contrast, electronic formats change frequently and require specific read/write equipment, technology that can become lost on more epochal time scales [135, 139]. This molecular conservatism will help drive recovery irrespective of what version of DNA sequencing technology is available even in the far future. Taken together, these features highlight the possibilities for DNA for extraordinarily long term archiving.

DNA-based information storage has previously been explored (Table 5.6), as challenges with conventional digital storage methods became

apparent. Bancroft et al. 2001 provided an early proof of concept and laid out the theoretical framework for encoding data into DNA. They proposed an encoding scheme that maps triplet combinations of the three nucleotides A, C, T to ternary numbers and uppercase letters of the English alphabet, along with a space character. The information is subdivided into segments (called iDNAs), and each segment is prefixed with a spacer and unique primer sequence, and then flanked with universal forward and reverse primers. An ordered array of unique primers is also included as a "polyprimer key", also flanked by the universal primer. With this, they encode the famous opening of A Tale of Two Cities by Charles Dickens in two iDNAs 232 and 247 bp long. The iDNAs can then be individually recovered by PCR and sequenced. Notably, this conception predates the revolution in reading DNA sequences wrought by NextGen sequencing technology, which we extensively leverage in our own work.

Church and colleagues reported in 2012 the storage and subsequent recovery of a 5.27 megabit stream of information (composed of text, images and source code)[125]. The information was encoded with a degenerate code, mapping A or C to 0 and G or T to 1 (this provided some freedom in preventing homopolymer runs and GC-skew). The information is partitioned into 54,898 blocks 159 bp each (which would amount to 8.7 Mbp of DNA), of which only 96 bp encodes information while 19 nt is used for address blocks and 44 bp for universal primers (incidentally this implies a rate of about 0.3 input bits per output bit, very close to that which we calculate for our encoding strategy).

After sequencing and filtering for perfect reads, they obtain 3000-fold coverage of each piece (for comparison, our simulations demonstrate data recovery with 5-fold and 10-fold sequencing). However, their demonstration still contained 22 nucleotide errors, of which 10 were also bit errors. Interestingly, this work notes that future work could incorporate compression, redundant encodings, parity checks and error correction - three of which are central aspects of our own work.

Goldman and colleagues in 2013 describe the encoding and reconstruction of various computer files totaling 739 kb[126]. The digital information was first Huffman coded (which reduces the space taken up by frequently occurring symbols) and then converted to a ternary representation, which is encoded into DNA at one nucleotide per ternary digit, rotating the nucleotides to prevent sequences that are difficult to synthesize or sequence from occurring. This single stream of DNA is broken into 100 bp fragments with 75 bp overlaps similar to our approach, but with alternate forward and reverse complement information in subsequent pieces. The overlaps are used only for redundancy, not assembly, and instead an address block is added to each piece to enable assembly. Their final encoding produces 153,335 DNA sequences that are each 117 bp long (totaling 17 Mbp), including the address information, which corresponds to a rate of roughly 0.17 (including the four-fold overlaps). After sequencing and filtering, they obtained a set of reads with 1308x coverage, from which they were able to recover the encoded digital data after applying error correction. For comparison, in our case we simulated read depths of 5x

and 10x, after which all of the data was recovered perfectly without need for error correction, with the exception of a few cases in which a single terminal nucleotide was lost. These losses could nonetheless be rectified by applying error correction so that the original data was recovered intact.

A 2015 publication by Grass and colleagues demonstrates encoding of 83 kB of data in DNA, using an error correction approach based on Reed-Solomon codes[140]. In this implementation, bytes of input information and block indices are mapped to elements of a Galois Field with an addition of Reed-Solomon redundancy for error correction. The Galois numbers are in turn mapped to 3 bp "codons" of DNA. Some codons are excluded from this mapping, which ensures that homodimers in output sequence cannot be longer than 3 bp. The output sequence is synthesized as 117 bp long pieces of DNA and stored encapsulated in gel. From 83 kb of text, they produced via this method 4991 pieces of DNA, for a total of 584 kbp of encoded DNA. Thus their encoding rate was roughly 0.59. Interestingly, Grass et al. performed experiments simulating thermal damage to the DNA in order to assess their error correction ability, and were able to recover original information exactly from a simulated 1% rate of per-nucleotide error; in our own simulations we also observed a threshold of about 1% for successful recovery. However, our simulations of error recovery did not account for error correction by the overlapping packets, and consider only errors that are not eliminated during the sequencing and assembly stages, so in reality our simulations underestimate the error-correction capacity of our work (a more detailed

discussion of estimated error rates is included in the Additional file 1).

Another recent report by Yazdi et al.[127] has made several advances to address-based encoding schemes by demonstrating a rewritable storage system which also supports random access. They distributed a total of 17 kB of digital information across 32 1 kbp gBlocks produced by a DNA synthesis technology. Using larger fragments allows for more efficient storage and reduces likelihood of data loss due to missing packets, both with address-based strategies as well as in our case. However, synthesizing of large amounts of DNA in bulk can be more economical with array based synthesis. The authors note that the cost of DNA synthesis is rapidly decreasing as new technologies are developed; should gBlocks or another method for synthesizing longer DNA pieces become cost effective in the future, our encoding scheme would be directly compatible with it. In fact, gBlock sequences are subject to a number of constraints, including lack of high GC areas, repetitive stretches and toxic sequence. Our strategy directly mitigates all of these issues as is, and is therefore fully compatible with gBlocks. Yazdi and colleagues also demonstrate random access and editing, by using specialized address sequences and PCR. Random access and ability to rewrite is a strength of address-based approaches – though limited random access could be added to our scheme (by flanking each packet with unique primer binding sequences, for instance) rewriting in place is likely to be impractical. Thus, we consider address-free approaches such as ours to be more suited for very long term, infrequently accessed, "cold storage" for archival purposes, where the latency and cost of editing is less

significant. Address-based systems, especially with rewrite capability such as that shown by Yazdi et al., would be well suited to a more frequently accessed storage tier above that utilizing ours.

Our strategy represents several important innovations over previous work. Most importantly, we do not rely on address blocks to guide assembly of the DNA sequence for decoding, but instead make use of the overlaps between packets of data. These overlaps can be used to construct a De Bruijn graph and be assembled in a manner analogous to the well-studied problem of genome sequencing and assembly. A primary advantage of this method is that only standard assembly algorithms, not specialized software, are required to perform the assembly. In addition, the different parts of the packet are equally vulnerable to mutation. With an address scheme, mutations falling on the address section can lead to the loss of the whole block, while mutations falling on the data portion are far more limited in scope. In our approach, if the packets tile the original sequence uniformly, it does not matter where the mutations fall, provided they are sufficiently sparse. This greatly simplifies the design of a single universal error correction scheme. Finally, the overlaps act as additional mutation buffers. Our encoding has tunable redundancy that already confers an error-correction capacity, but the added redundancy from overlaps allows straightforward detection and correction of sequencing errors and rare mutations as a result of the de novo assembly process. This "failsafe" error correction protocol makes this strategy the most secure thus far for any long term storage attempts.

With address blocks, an important consideration is efficiency of storage: within each packet, the address information and the encoded data compete for space. The number of packets scales with the total size of the encoded digital data, and the associated address block becomes larger when there are ever more packets. An upper bound on the number of packets is the number of possible unique sequences of that length (in reality this tends to be a gross underestimation since many possible sequences are unsuitable for synthesis and sequencing). Early on it was recognized by Goldman and colleagues, for instance, that with 114 bp packets, 14 bp could be used for indexing to obtain 88% efficiency; this imposes an upper limit of about 268 million packets covering 6.7 Gbp of sequence (taking into account the four-fold redundancy). For a similar level of redundancy, our approach could generate 114 bp packets tiling the sequence with 85 bp overlaps, for a total of 231 million packets. The upper limit on how much information can be stored would most likely be determined by the limits of assembly software. Moreover, arbitrarily large amounts of information can be stored regardless of this upper limit, if separate pools of encoded DNA are physically isolated by storing them in their own individual containers. Alternatively, multiplexing is possible by flanking packets pertaining to each encoded file with a unique pair of primer sequences, and then selectively amplifying that subpool as needed (this also allows for random access[127]).

Our overlap-based approach therefore allows for a simpler and more straightforward partitioning of encoded data than relying on address blocks. It

eliminates the need for parsing and interpreting address blocks when decoding, as well as the concern over how to preserve the instructions for performing this step correctly. The sizes of address blocks must be explicitly standardized prior to encoding, which will impose an upper limit on how much data can be encoded for as long as that standard is in effect. If the address blocks are too small, large encodings will run out of address space and the encoding standard will have to be changed often (leading to compatibility issues), but if they are too big then small pieces of data will be encoded inefficiently. Our approach sidesteps this dilemma; it is only necessary that sufficient overlap exists between packets to enable assembly, and no space is "wasted" on address blocks since the overlaps act as an additional safeguard of data integrity. The information is partitioned efficiently regardless of its size, the upper bound on capacity is very large, and no specific details of the partitioning need be recorded, since simply attempting to assemble the overlapping pieces of DNA (a step which would be obvious even if the encoding scheme is unknown) yields the complete stream of encoded data.

5.5 Conclusions

We have described the design and in silico implementation of a strategy for encoding digital information into distributed DNA sequences and then reassembling the original information irrespective of intervening errors. We evaluated several key parameters, including the manner in which the sequence would be partitioned into oligonucleotides, the cloning regimen necessary to maintain a sufficient fraction of the original information for reconstruction following dispersal, the read depth needed to ensure successful recovery, and limits on how many mutations can be tolerated while still allowing complete recovery of the information. Our simulations demonstrate a successful round-trip of information during long term storage in DNA, taking into account problems that might be encountered during synthesis (the write phase) and sequencing (the read phase).

The error correction method we have chosen, and the closely related algorithm for codebook generation, facilitate tailoring the encoding to specific applications. By reducing the minimum distance between code words, one can reduce redundancy, thereby allowing the encoded data to fit into shorter DNA. Conversely, when the integrity of the data is more important than the efficiency of storage, the minimum distance can be set very high, thereby potentially allowing recovery even if a large number of mutations are introduced.

By compressing our input data with LZMA, we remove correlations between the statistical properties of the input data and those of the resulting DNA sequence. LZMA is a compression algorithm based on constructing

a Markov model of the data that captures recurring patterns within it, and then encoding the overall data in terms of this Markov model. To the extent that patterns do exist, the resulting information is compressed and largely free of patterns, resembling random data or uniform digital noise. Thus, LZMA should allow us to produce uniform and non-repetitive DNA sequence from even very skewed and / or repetitive inputs. By explicitly generating a codebook with well-defined parameters, we should be able to further impose constraints on any resulting sequence output, such as requiring a GC composition that would be optimal for synthesis and sequencing, irrespective of the composition or structure of the input data.

In our work we used a codebook of eight 4 bp blocks. In principle, it is possible to obtain the same level of overall redundancy with longer blocks, the minimum codeword distance being scaled up as appropriate. Two key parameters for choosing the block size were performance (codewords for longer blocks take more time to generate) and the expected distribution of mutations. Longer codewords, with more distance between them, have a greater per block mutation tolerance, even if the per base pair mutation tolerance remains the same. Error correction functions only if the number of mutations within a particular block remains below a threshold; that is, mutations must not be too close to each other. If there is a tendency for mutations to occur in small clusters, rather than being evenly distributed, longer blocks would likely be preferable.

Though our primary interest was to develop the means of storing infor-

mation in DNA, the process can be adapted to a number of other applications. Herein we will discuss three of these: repurposing the codebook generator to produce barcodes, using encoding for inactivation of toxic sequences and the potential to use DNA-encoding to protect private information.

We designed our codebook generation algorithm to produce a set of sequences that are evenly spaced in sequence space. This encoding has similarities to the mathematical concept of a "perfect code"[141] which refers to sets of codewords optimally arranged in sequence space so as to provide a given level of error correction, although our approach considers Levenshtein distances rather than Hamming distances, and furthermore there are codewords we deliberately avoid, so in practice our code is less than perfect. Though our purpose was to encode information, an interesting property of the resulting codewords is to lack bias and repetition, and therefore be distinct from one another. The relatively small length makes them easy to synthesize and the lack of repetition makes them easy to sequence. Overall, these properties also potentially make them excellent DNA barcodes[143, 142, 144], and as such they can be readily applied to a long list of biological techniques such as multiplexed NextGen sequencing[145], identification of genetic variants[146], construction of deletion libraries[147] and high-throughput RNA profiling[148].

Another consequence of encoding is that output sequence has very little resemblance to input. This can make the storage of difficult sequences more tractable. For example, if a given DNA sequence is fragile or unusually prone to mutation, encoding should allow it to be stored with high fidelity. Our cen-

tromere simulations illustrate this principle - although centromeric sequence is not per se toxic, it is very repetitive, impairing its faithful construction and sequencing. In contrast, when encoded, the centromere data does not have any obvious repeats or other unwieldy sequences. Encoding can be used to not only preserve information, but also to prevent its ready expression in the absence of decryption. Any biological sequence is rendered biologically unreadable by our scheme for encoding, and this in turn suggests that biothreats such as toxin genes or even entire pathogen genomes (e.g. genomes of infectious viruses) that would otherwise be harmful in their 'biologically readable' form could potentially be encoded and stored over long periods, essentially placing them in 'deep storage' for posterity.

Distinctly from previous work, we encode information such that there is a specific codebook associated with each piece of information, which is also necessary for later decoding. In this sense, our encoding method is analogous to a symmetric key cryptography scheme, with the codebook as the key. With short block lengths, it is conceivable that the codebook might be reverse engineered from the encoded sequence only by an exhaustive approach. With longer block lengths, it would become extremely difficult to recover the information without access to the codebook. This allows information to be hidden from readers who do not possess the correct codebook, or simply to prevent information from being accessed before an arbitrary amount of time has passed (that is, the time needed for exhaustive reconstruction of the codebook). These features render DNA encoding inaccessible and resilient in a way

that far exceeds what can be availed via electronic encoding. While there may be few applications for information privacy that can tolerate slower read-write capabilities, it should be noted that mechanical, Enigma-like rotor cypher systems can be difficult and time-consuming to crack even with modern electronic computers [26]. Thus, it is entirely possible that a nation state or corporate enterprise might wish to hide (and retrieve on a leisurely scale) 'trade secrets' via the long term, difficult to break, and undegradable encoding strategies we describe herein.

Finally, because of the Markov chain-based compression employed, having access to part of the encoded DNA does not allow one to access part of the original information. Decoding is only practical if the entire sequence is present, since otherwise both assembly and decompression fail. This all-or-nothing nature of the information round trip requires care to be taken that a sufficient number of packets are retained throughout storage; in our case we solve this by creating a redundant set of packets that allow assembly even if a small number is lost. Subsets of the packets could also be distributed to different recipients or locales, which would then be unable to decode the information on their own without combining their respective pieces of the DNA-encoded data. This strong message compartmentalization, combined with the potential for multi-hundred thousand year digital file storage capacities–a time-scale exceeding all known continuous human civilizations–allows one to consider the prospect of securely preserving information such that it outlasts even extreme, global catastrophic events.

To summarize, we have demonstrated a strategy for encoding digital information into DNA, which is highly parallelizable and has built-in error correction ability, and relies on an overlap-based assembly method fundamentally different from the previously published approaches. We have demonstrated full round trip read and write of the information, including simulated sequencing of a cloned oligo pool, with recovery from simulated mutations.

5.6 Methods

5.6.1 Test data

The Cat.jpg file was obtained by scaling down a color photo of a cat (released into the public domain) and encoding with the Jpeg algorithm. After adjusting the pixel size of the image to approximate our desired size, we then fine-tuned the Jpeg compression level (ultimately we used level 86) so that the resulting file size was about 10 kB. Cat-big.jpg was constructed similarly to produce a 100 kB image. The random data file was obtained by creating a 10 kB file and populating it with random data piped from /dev/urandom on a Unix computer. The centromere input was a text file containing part of the sequence from human chromosome I (NCBI sequence NC_000001.11 positions 11,000–22,000). The flat file was constructed by pasting a large number of 0 s into a text file. In every case, the test data was wrapped in a tar archive (which is a file format that can store one or more files without compression), which is then used as the input for the DNA codec.

5.6.2 Codebook generation

The codebook is a table showing permissible blocks of DNA sequence, and what numeric value each sequence maps to. Our algorithm (implemented as Python code in the file make_codewords.py) begins by generating all possible sequences of a given length (Table 1). This pool is filtered to remove sequences with high repetitiveness (quantified by dividing the number of its unique subsequences to the number of all of its subsequences, implemented in

complexity_estimation.py) and undesirable GC content (in our case set as less than 40% or more than 60% GC). Of the remaining sequences, one is picked at random and saved as a codeword. All sequences with Levenshtein distance less than the defined threshold (in our case 3) are removed. Of the remaining sequences, another one is picked at random to be the second codeword, those with too small Levenshtein distance are pruned again, and the process is repeated until no further codewords can be produced. Sequential integer values starting from 0 are then assigned at random to each codeword in the resulting codebook. We generated a single codebook of eight 4 bp codewords and used it for all of our experiments (Table 2).

5.6.3 DNA codec implementation

The DNA encoding-decoding scheme was implemented in Python (dna_read.py and dna_write.py) to produce a codec which, given the codebook, converts a digital file into a DNA sequence (encoding) and vice versa (decoding). For encoding, the file is first read, compressed with LZMA (lzma library of Python 3.5.2), converted into a byte array, which is equivalent to the digits of a base 256 number. A base change operation is performed, to change the base from 256 (the number of possible values a byte may have) to 8 (the number of codewords in our codebook). The resulting array of integers is converted into short DNA blocks according to the mapping in the codebook, and concatenated into a single DNA sequence. To decode, the entire DNA sequence is broken into 4 bp blocks, then each block is mapped

back to an integer according to the codebook, producing a base 8 number. A base change is performed from 8 to 256, and the resulting array of numbers is converted into a byte stream and decompressed with LZMA.

5.6.4 Error correction

The error correction is applied by scanning through a DNA sequence and checking each subsequent block against the codebook. If the block appears in the codebook, then it is deemed correct. Otherwise, the block is deemed to be mutated, and the Levenshtein distance from the observed block to each code word is calculated. The block with the minimum distance is assumed to be the original sequence before mutation. This was implemented in clean_dna.py and related files.

5.6.5 Partition of encoded DNA into packets

In order to accommodate eventual synthesis of the encoded DNA sequence into nucleic acid, it is broken up into short subsequences termed "packets" (implemented in split_into_packets.py). The packets are selected so that each one is 200 bp long (selected because it is the longest length feasible with current large-scale DNA synthesis technology), and they tile original sequence such that each two adjacent packets overlap by a specified amount. For the beginning and end of the sequence, shorter packets are produced to avoid lower coverage of sequence ends.

5.6.6 Sequencing simulations

We simulated NextGen sequencing of our encoded DNA and packets using the ART sequencing simulator[149]. We simulated 150 bp-long reads with the Illumina HiSeq 2500, this being a common, commercially available method of sequencing. The read depth varied from 1x to 50x depending on the particular experiment.

5.6.7 De novo assembly

Fastq files generated by the ART simulator were used for de novo contig assembly with the Geneious assembler from the Geneious 9.1.2 software[96]. The parameters specified that the expected assembly was linear and not circular. The longest resulting contig was then taken as the assembled sequence.

5.6.8 Mutation simulation

To simulate mutations, we implemented a simple iterative algorithm (mutate_dna.py). In each iteration, the algorithm picks a random base pair and changes it to a different, randomly picked base. This operation is repeated as many times as needed.

5.7 Supplementary Text

5.7.1 General estimates of error tolerance

In any round trip write-store-read experiment, we can group errors into three classes:

- Errors arising during synthesis, with per nucleotide rate $s' = 1 - s$
- Errors arising during storage, with per nucleotide rate $r' = 1 - r$
- Errors arising during sequencing, with per nucleotide rate $q' = 1 - q$

These groupings are advantageous in that the parameters of one can often be manipulated independently from the other two. For instance:

- Synthesis errors can be improved by using more sophisticated synthesis technology or producing shorter pieces of DNA
- Storage errors are influenced strongly by duration of storage, storage medium and temperature
- Sequencing errors depend on choice of sequencing technology and prep

For any given position in the encoded data, the fraction of DNA molecules in the synthesized pool that carry the correct nucleotide at that position will be s. After storage, this proportion will fall to $s \cdot r$. The proportion of sequencing reads covering this position that also correctly report the unaltered bases will be $p = s \cdot r \cdot q$.

5.7.1.1 Correcting errors by majority consensus

After processing sequencing data and obtaining c reads covering each position, the probability of a majority of reads reporting the correct base is given by the binomial distribution $P = B_{\mathrm{CDF}}(c, k, p')$ representing probability of no more than k successes from n trials with probability of success $p' = 1 - p$, where k is $(c + 1)/2$ rounded down.

Picking a modest estimate of $c = 5$, we get $k = 3$. In our simulations we have determined that redundancy-based error correction can recover from an error rate of up to $P = 0.01$ in the final assembled consensus sequence. This implies a maximum threshold for the round trip error rate $p' \approx 0.10$. If coverage is increased slightly to $c = 10$, then $k = 4$ and $p' \approx 0.15$. For much higher but still reasonable $c = 50$; $k = 24$ and $p' \approx 0.33$. Note, however, that this does not include the effect of errors on assembly.

5.7.1.2 Real world estimates of error rates

Synthesis errors tend to be on the order of 1% to 0.01%[150], although they are often strongly influenced by certain parameters such as desired fragment length (shorter pieces have less error) and yield (there is a trade-off between fidelity and total yield). We can take $s' = 0.01$ as an estimate easily achieved by even older technology. However it is clear that much higher synthesis fidelities can be achieved — for example, in our case encoded information is composed solely of a limited repertoire of n-mers, so synthesizing DNA from pre-manufactured n-mers instead of monomers would not be tech-

nically challenging and could improve the segment length several fold, albeit with the potential to introduce segmental substitution error.

The experiment of Grass et al.[140] can be interpreted as a measurement of r'. They report per nucleotide error rates on the order of 1% after simulated storage for 2000 years. Thus we can take $r' = 0.01$.

Sequencing error rates can vary substantially depending on the particular equipment and protocol used. Notably, statistical properties of the sequence (such as repetitiveness) also have a strong effect, and our encoding method minimizes problematic sequence features. Nevertheless, commonly reported rates are on the order of 0.1-1%[151] and from this we can take $q' = 0.01$.

Based on the above, we can estimate $p = s \cdot r \cdot q = 0.99 \cdot 0.99 \cdot 0.99 \approx 0.97$ and correspondingly $p' = 0.03$ for a reasonable expectation of the error rate in practice, after storage for many centuries. This is well below the maximum even for 5x coverage, which is $p' = 0.10$. Thus, high-fidelity storage appears to be very feasible using the method we propose, even on extremely long timescales and modest technical resources.

Note that for the above calculations provide an estimate of per-nucleotide errors, but it is difficult to precisely calculate true error rates in actual usage due to the many, sometimes unpredictable factors involved. For the above estimates, it is assumed that:

1. Assembly is successful. The likelihood of this is explored in the sections

of our main manuscript titled "Library construction and long-term packet-wise retention" and "Simulated recovery of information with packet loss".

2. There is no block-wise synthesis error (as would be the case if packets were synthesized from n-mers rather than individual monomers, for instance).
3. There are no systematic errors or jackpot effects. Our main manuscript, in the section "Encoding of digital data into DNA", discusses the features of our codec which enable mitigation of systematic errors.

Therefore, real world performance will likely be lower than these estimates.

5.7.2 Details of Data in Table 5.6

5.7.2.1 Bancroft 2001

The phrase encoded was "IT WAS THE BEST OF TIMES IT WAS THE WORST OF TIMES IT WAS THE AGE OF FOOLISHNESS IT WAS THE EPOCH OF BELIEF"[124], which is 106 characters long. The authors specify that they mapped values to uppercase letters and a space, a total of 27 possible characters. In this case, Shannon information would be $107. \log_2 27 = 509$ bits.

5.7.2.2 Church 2012

The publication reported encoding of 5.27 million bits in total. This was stored on 54,898 sequences, each 159 bp long, for a total of 8,728,782 bp. [125]

5.7.2.3 Goldman 2013

The numbers for total bits of information and total bases used were taken as reported in the publication. [126]

5.7.2.4 Grass 2015

It was reported that 83 kbytes of information was encoded ($83,000 \cdot 8 = 664,000$ bits) and 4,991 segments of DNA, each 158 long, were produced ($4,991 \cdot 158 = 788,578$ bp of DNA).[140]

5.7.2.5 Yazdi 2015

While the total raw input information was given as 17 kbytes, this study encoded information using a dictionary of words rather than traditional byte-characters. We have therefore taken their own figure of 23,196 bits of information for the word-based encoding (Table S1).[127]

It was reported that the total of bases produce was 32 kbp.

5.7.2.6 This publication

The Cat.jpg image was 10478 bytes, or $8 \cdot 10,478 = 83,824$ bits. The encoded DNA produced from this was 111,192 bp long.

5.8 Abbreviations

cpo Clones per unique oligonucleotide

GC Guanine-cytosine (composition of DNA)

Jpeg Joint Photographic Experts Group (image compression standard and the file format implementing it)

LZMA Lempel–Ziv–Markov chain algorithm (compression algorithm)

oligo Oligonucleotide

ORF Open reading frame

5.9 Acknowledgements

We are immensely grateful to Randall Hughes for invaluable discussions regarding the technical details of DNA synthesis technology and its limitations. EMM acknowledges grant support from the NIH, NSF, CPRIT, and Welch Foundation (F-1515).

5.9.1 Funding

EMM acknowledges grant support from the NIH (R21 GM119021, R01 HD085901, DP1 GM106408, R01 DK110520, R35 GM122480), NSF, CPRIT (Cancer Prevention and Research Institute of Texas), and Welch Foundation (F-1515). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

5.9.2 Availability of data and materials

The source code for programs described in this manuscript can be found in a public repository at . Data files and DNA sequences encoded and analyzed in this manuscript are also provided in the repository as example input and output data.

5.10 Authors' contributions

AA and EMM conducted the research. AA implemented the software, ran simulations and analyzed results. AA, ADE and EMM conceived of the main idea, and also wrote and revised the manuscript. All authors read and approved the final manuscript.

5.11 Notes

5.11.1 Ethics approval and consent to participate

Not applicable.

5.11.2 Consent for publication

Not applicable.

5.11.3 Competing interests

The authors declare that they have no competing interests.

5.11.4 Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

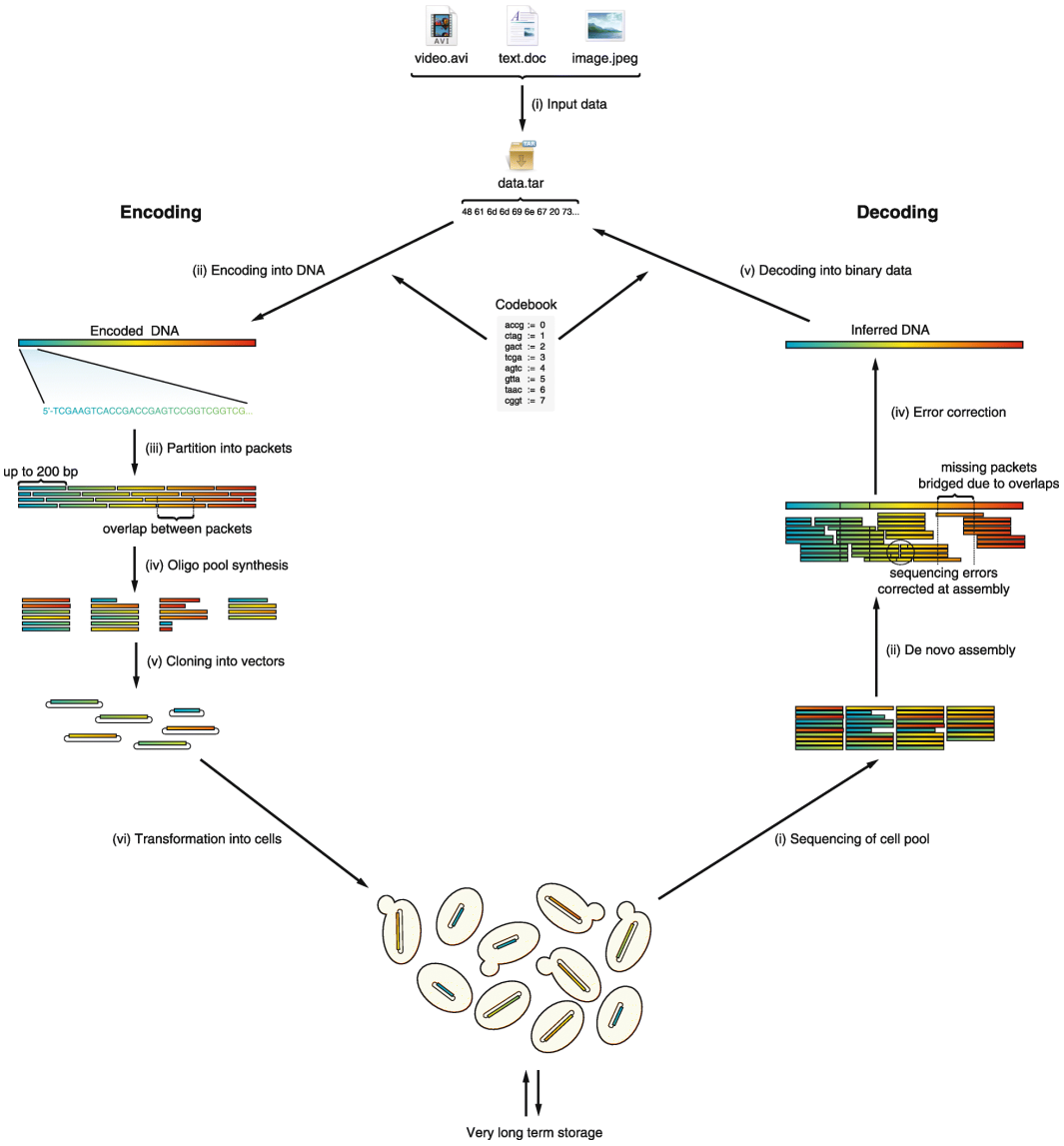

Figure 5.1: A diagram of the encoding and decoding process. (Continued on next page.)

Figure 5.1: The input data is first wrapped in a tar archive, to ensure a uniform input format as well as combining multiple files into a single contiguous data stream with a well-established method. The digital information is encoded using a pre-generated codebook, producing one long single sequence of DNA. This sequence is split into overlapping packets, each up to 200 bp long, which are then synthesized as a complex pool of oligonucleotides. These can be cloned into plasmids and transformed into cells, where they can be maintained reliably for a very long time. To recover the information, the population of cells (or alternatively plasmids or lyophilized oligonucleotides) can be sequenced with NextGen sequencing technology, and de novo assembly of the resulting reads is performed. During assembly, some errors can be corrected by simply considering the consensus of the contig, whereas systematic errors (such as those arising during synthesis) can be corrected in silico using the error correcting code. Finally, the codebook is used to decode the resulting contig and recover the digital files.

Figure 5.2: Digital data used for in silico experiments. A 300 x 200 pixel color photo of a cat, encoded with the Jpeg algorithm so as to produce a files 10,387 byte file.

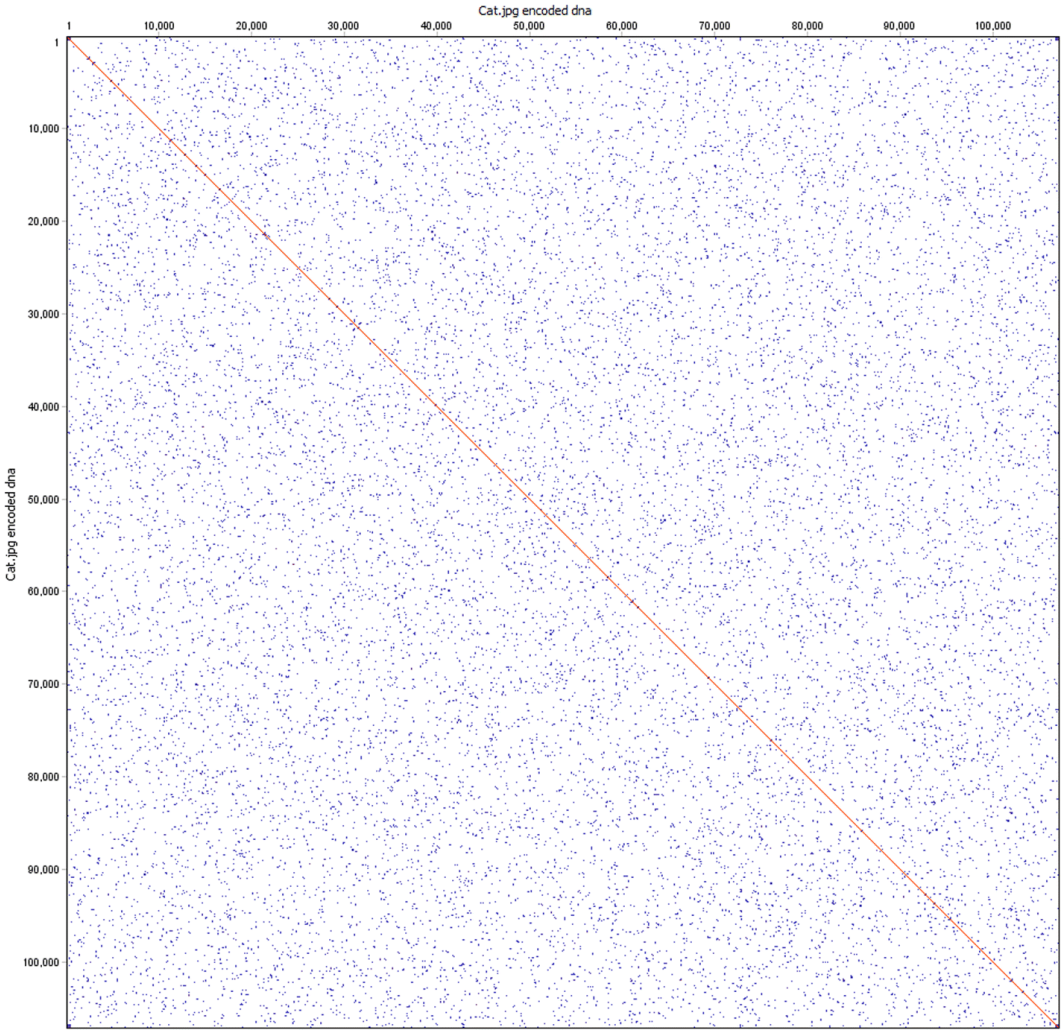

Figure 5.3: Overall self-similarity of the encoded Hamming image. Dot plot of the encoded Hamming image generated with dottup, using word size 20 as the parameter. Positions where 20 bp of the sequence are self-similar are marked with blue. Identical regions longer than 100 bp are marked with red. The plot shows a lack of long stretches of repetition that could interfere with assembly.

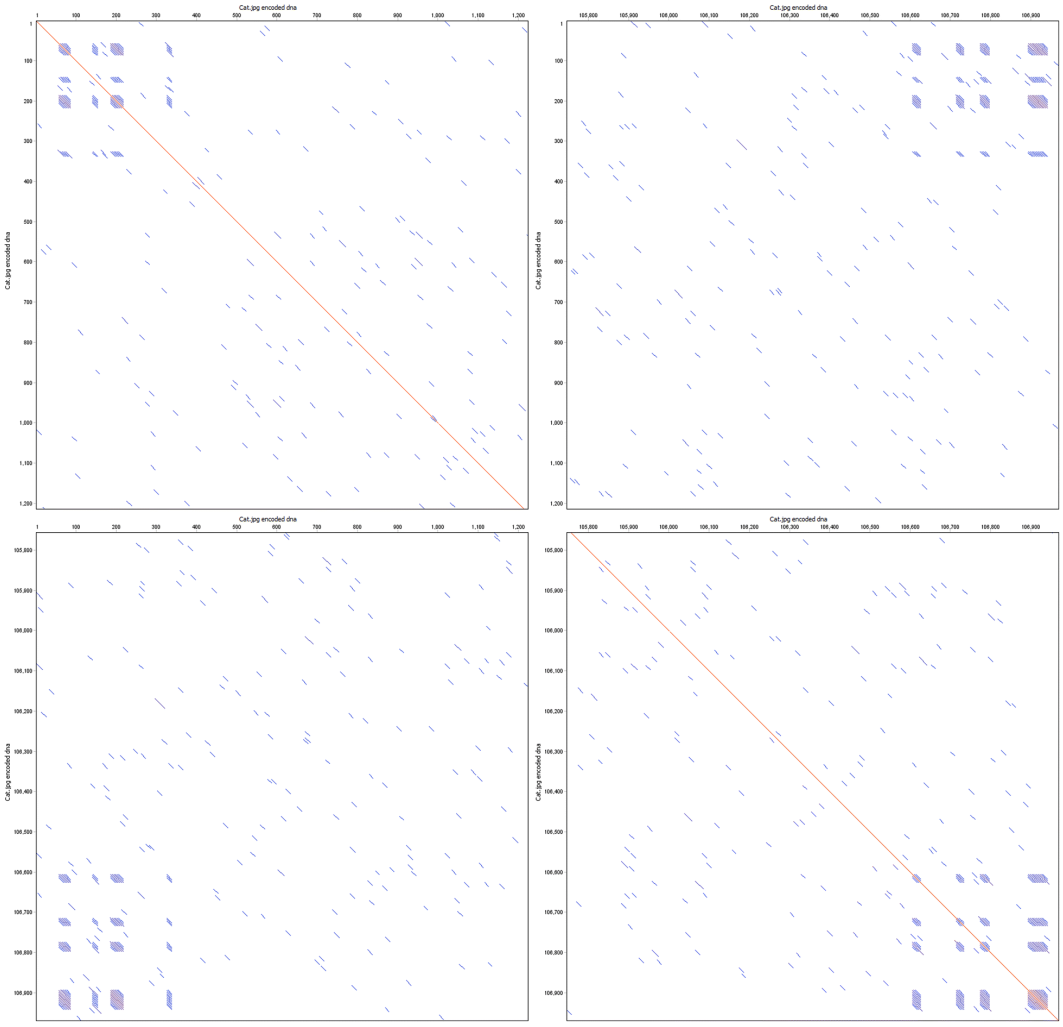

Figure 5.4: Self-similarity at corners. Dotplots of the same encoded DNA, showing only the ends of the sequence, generated with word size 10. Short blocks of repetitive sequence are visible as blue blocks, these result from header and terminator information utilized by the LZMA algorithm which is less variable than the compressed data stream itself. Top left: Sequence head vs. itself. Top right and bottom left: Head vs. tail. Bottom right: Sequence tail vs. itself.

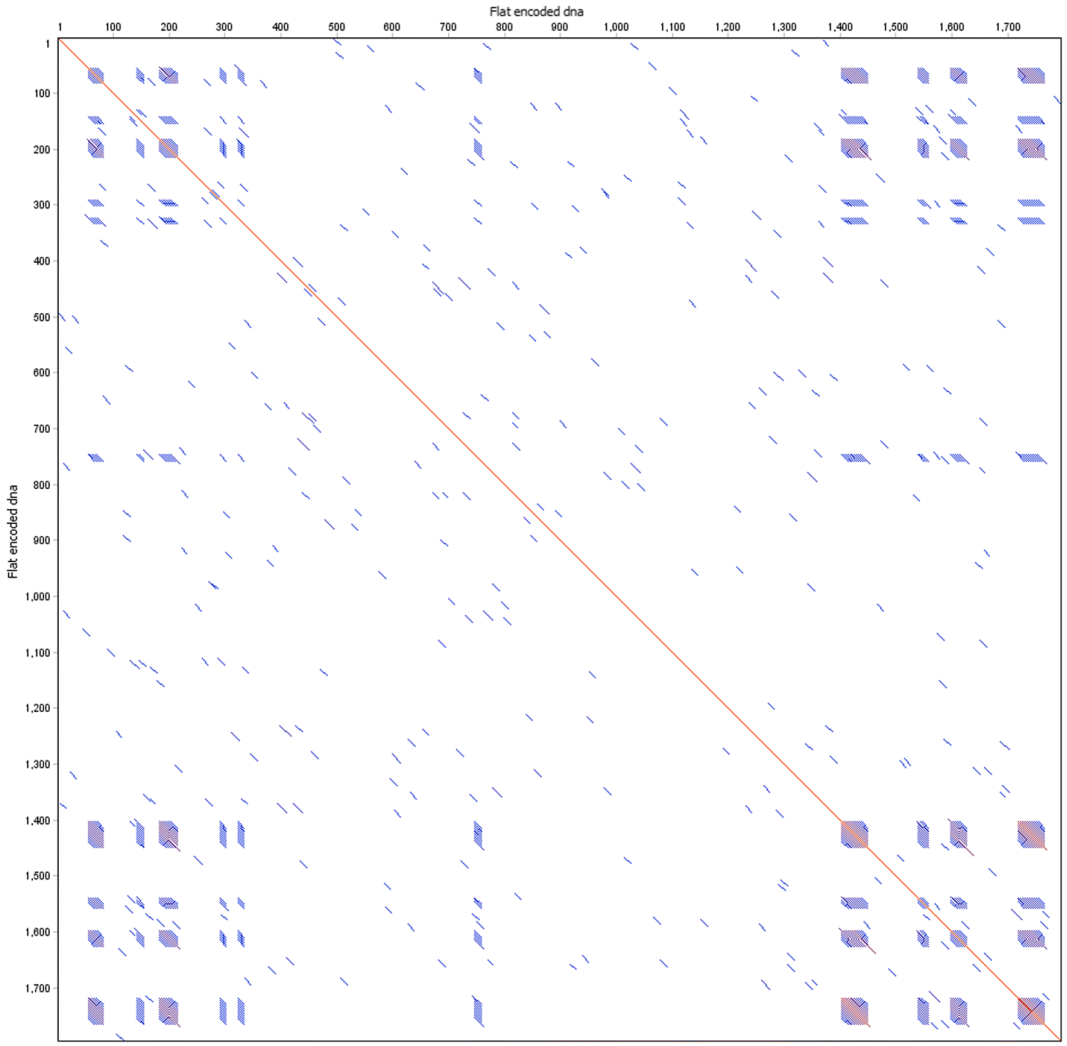

Figure 5.5: Self similarity of flat file. Dot plot of the entire encoded flat file, generated with dottup with word size 10, showing self-similarity within the entire encoded DNA sequence.

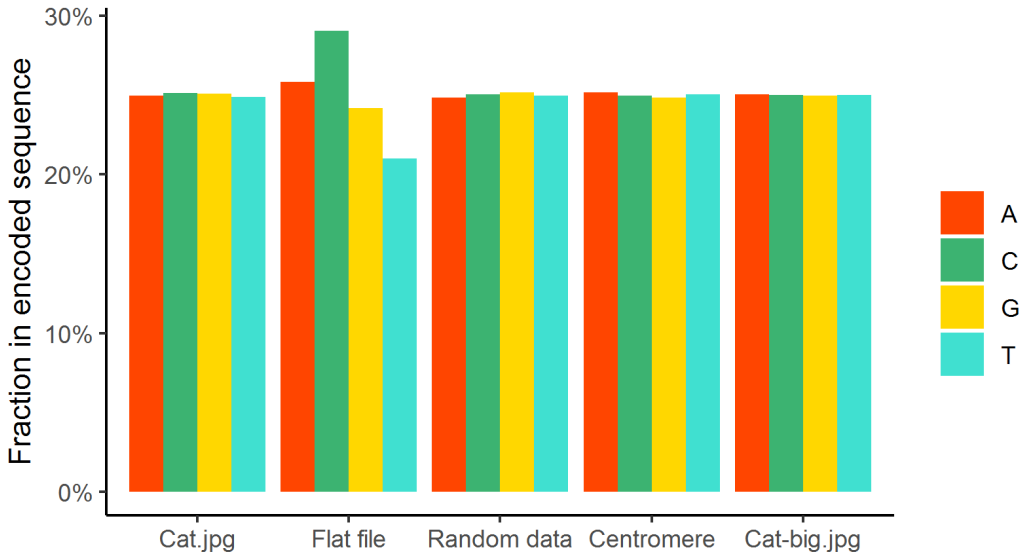

Figure 5.6: Total nucleotide composition of encoded DNA. Bars show the relative fraction of each nucleotide within DNA obtained by encoding the given digital data.

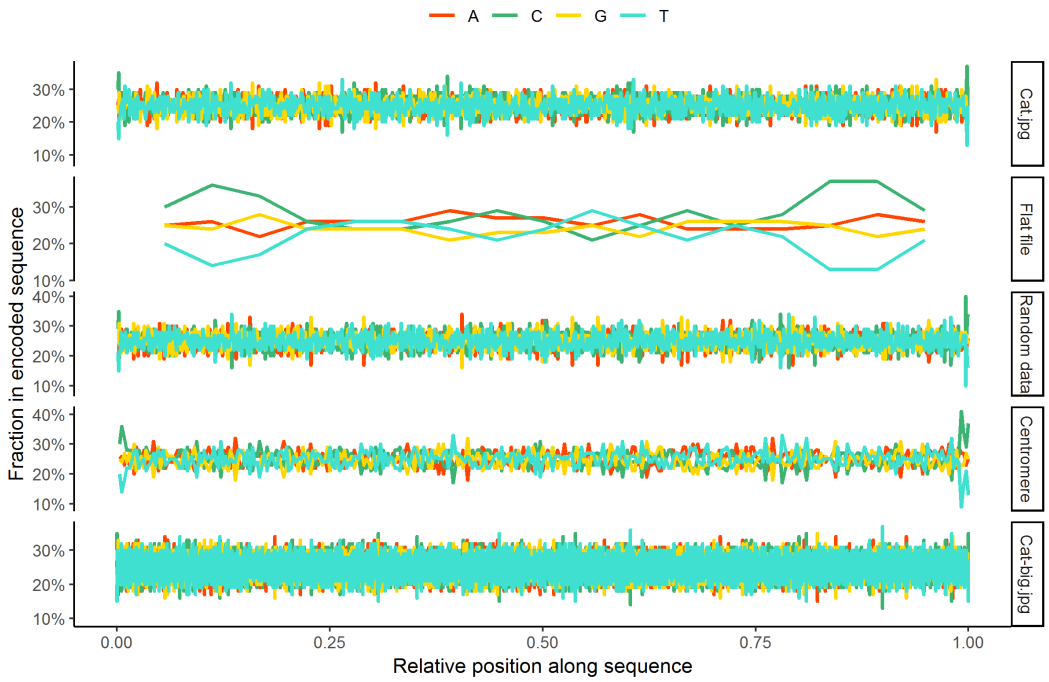

Figure 5.7: Local composition. Nucleotide composition in sliding 100 bp window for each sequence of encoded DNA.

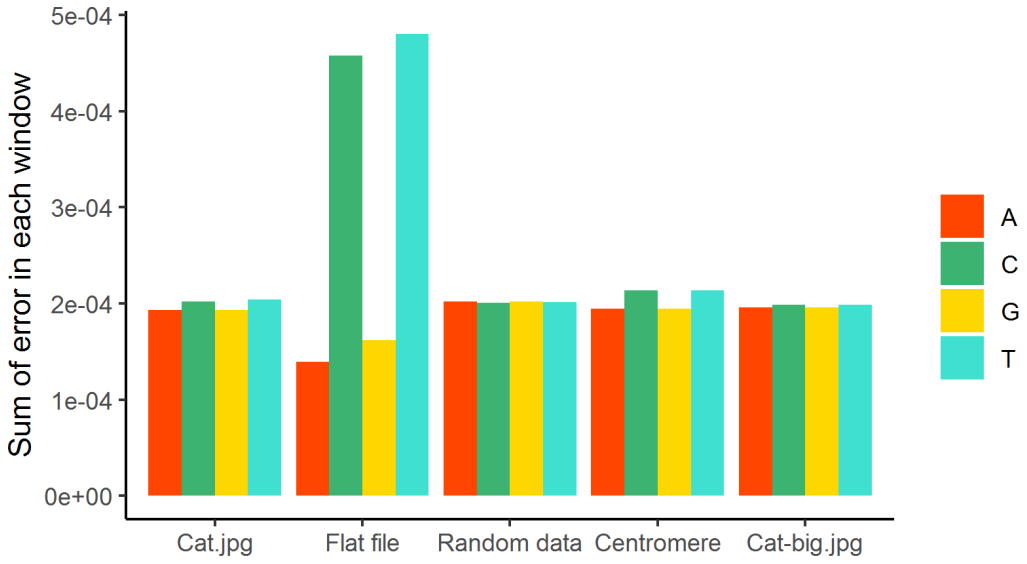

Figure 5.8: Total nucleotide composition error. Total deviation of nucleotide composition from the expected 25% proportion. Shown here is sum of error within each 100 bp window tiled along the encoded sequence, and divided by the sequence length.

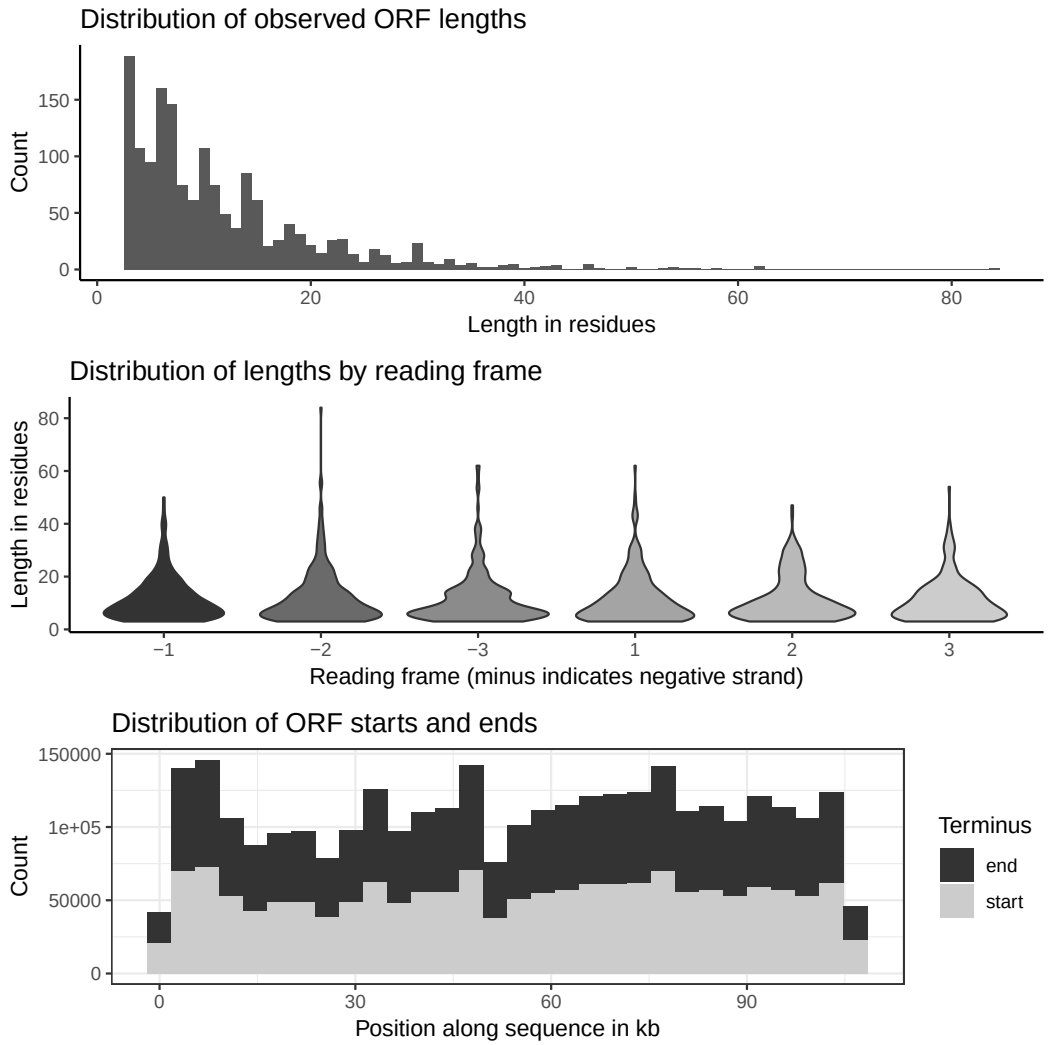

Figure 5.9: Spurious ORFs in encoded sequence. Top: Histogram showing the distribution of spurious ORFs observed in the DNA sequence for the encoded Hamming image. Middle: Violin plot showing the length distribution of spurious ORFs grouped by reading frame. Frames are marked with a minus (-) if they are on the negative strand (ie. detected in the reverse complement of the sequence). Bottom: Distribution of spurious ORF start (grey) and stop (black) positions along the sequence.

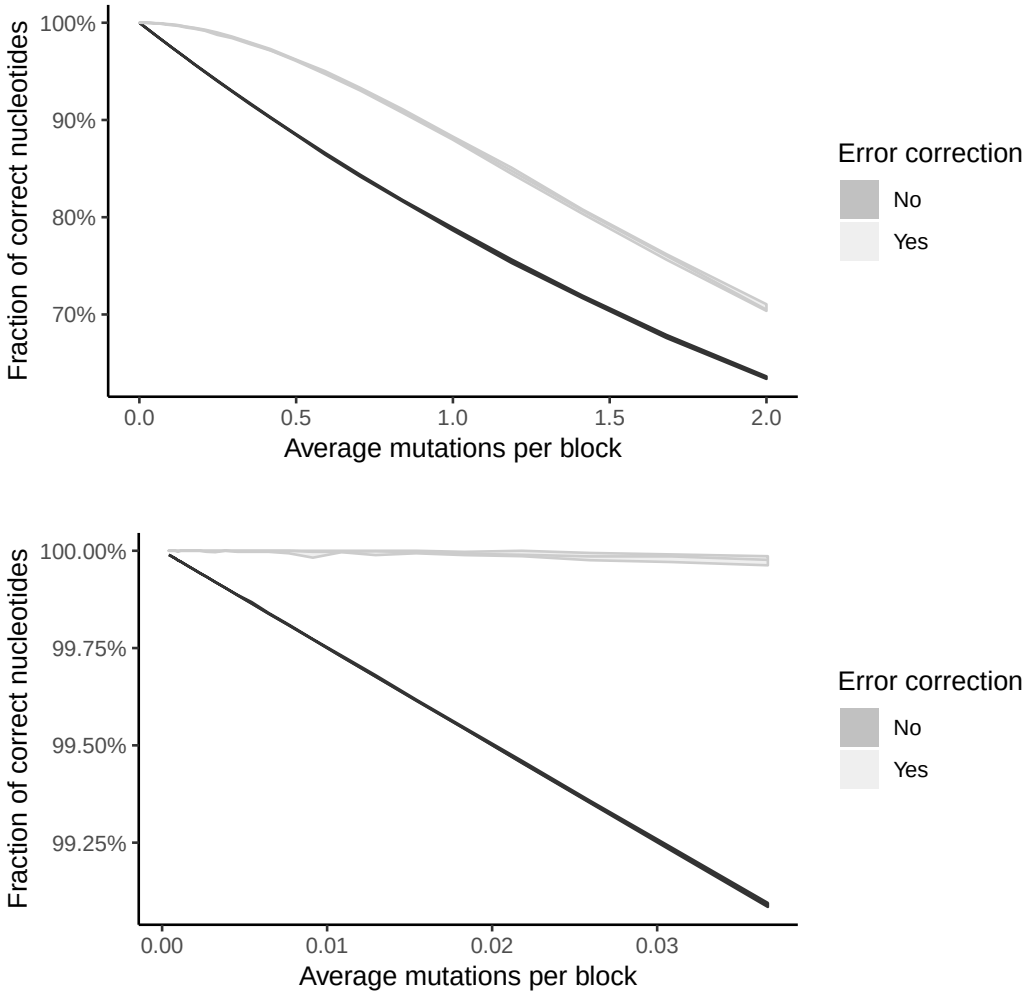

Figure 5.10: Mutation buffering by error correcting code. Light gray line shows the effect of applying error correction on sequences mutated to varying degrees. The mutation rate is shown in average mutations per block, given that each block is 4 bp long. Shown here is the mean of 10 simulations (middle line) with $\pm 2.58\sigma$ band (shaded area), representing 99% confidence interval. Bottom plot: Subset of the data corresponding to only lower mutation rates.

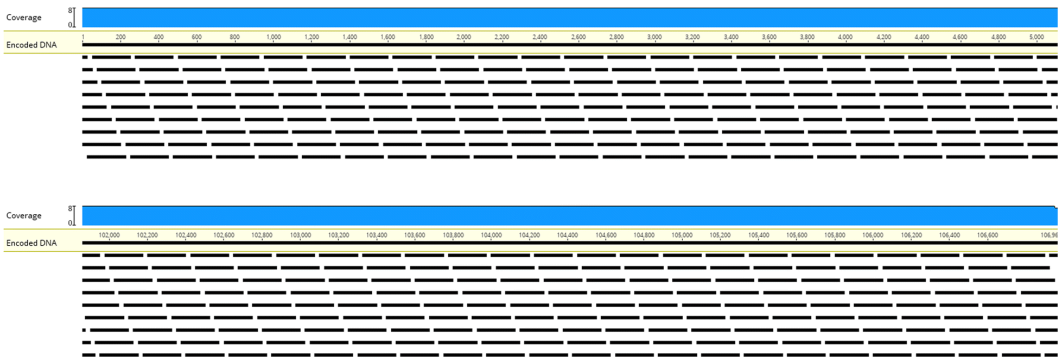

Figure 5.11: Distribution of oligos along the encoded DNA. Black bars indicate 200 bp oligos, produced so as to tile the encoded sequence with 175 bp overlaps between two successive oligos. Blue graph shows coverage of the DNA by oligos (uniformly 8x virtually everywhere). Only the beginning and end of the sequence is shown here; but the parts shown here are representative of the entire sequence.

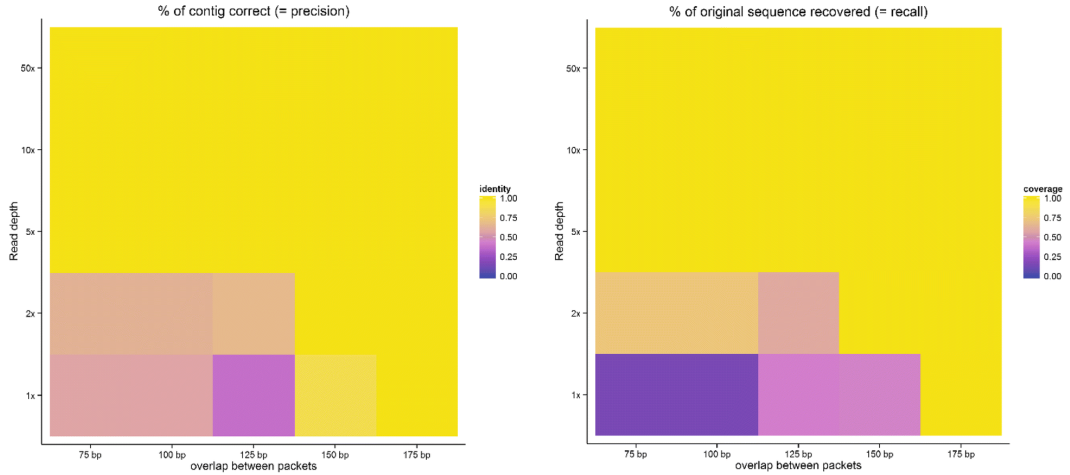

Figure 5.12: Left: Fraction of base pairs in the longest assembled contig that matched the original sequence after mapping. Right: Fraction of the original sequence that was present in the longest assembled contig.

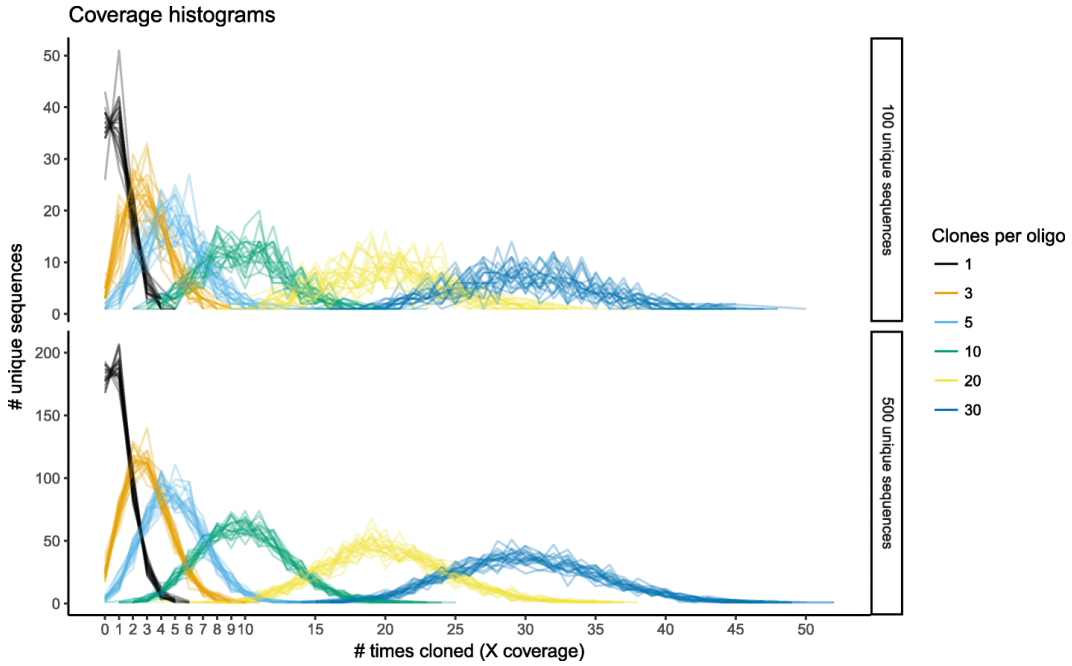

Figure 5.13: Monte Carlo simulations of library construction from pools of oligos. A series of simulated random draw experiments were performed for pools of 100 and 500 unique sequences; the number of draws ranged from 1 to 30 times the number of unique sequences. For each combination of parameters, 20 repeat experiments were performed, shown here as lines of identical color.

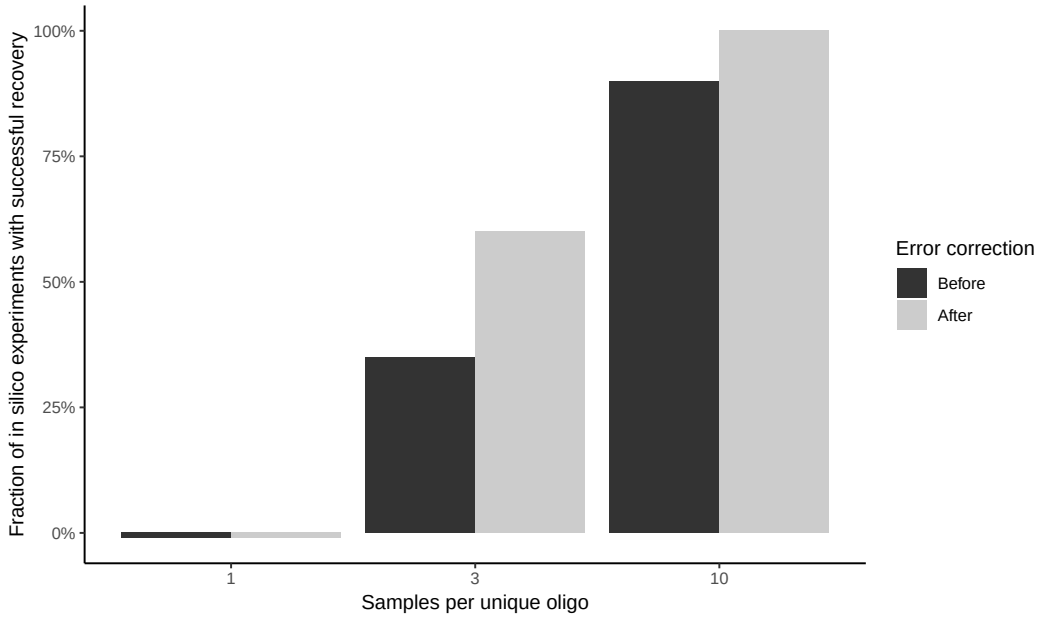

Figure 5.14: Recovery of original sequence after simulated sampling of packet pool and simulated sequencing. Each bar shows fraction of 20 simulated read-write experiments in which the data after decoding matched the encoded data exactly. For black bars, the error correction capacity built into the sequence was ignored, and the assembled contig was decoded as-is. For grey bars, the error correction was applied and decoding was attempted after.

Parameter	Explanation	Value
Word length	Length of each codeword to be generated, also determines cipher block size	4
Max mutations	Maximum number of mutations that can be recovered from per block. If this parameter is k, the algorithm generates codewords such that the minimum Levenshtein distance between them is $2k + 1$.	1
Min. GC	Minimum GC composition in each codeword	0.4
Max. GC	Maximum GC composition in each codeword	0.6
Complexity	Minimum complexity (used as a proxy for repetitiveness) of each codeword	0.75

Table 5.1: Parameters used for codebook generation

Codeword	Value
accg	0
ctag	1
gact	2
tcga	3
agtc	4
gtta	5
taac	6
cggt	7

Table 5.2: Codebook

Input data	Explanation	Rationale
Cat.jpg	Color photo in jpeg format, scaled down to 300 x 200 pixels	Example of real world data
Flat file	A text file containing a string of 10,340 zeros	Demonstrate performance when given extremely repetitive data
Random data	10 kb of random data obtained from /dev/urandom on a Linux computer	Demonstrate performance when given data without any statistical bias
Centromere	Part of centromeric sequence from human chromosome I, retrieved from: `https://www.ncbi.nlm.nih.gov/nuccore/NC_000001.11?report=fasta&log=seqview&format=text&from=11000&to=22000`	Real-world example of repetitive information
Cat-big.jpg	Higher resolution (1125 x 750 pixels) version of Cat.jpg	Example of larger (100 kB) input file

Table 5.3: Input data

Data	Size of file (bytes)	Size of tar archive (bytes)	Length of encoded DNA (bp)	Rate
Cat.jpg	10,387	12,288	106,968	0.459
Flat file	10,340	11,776	1,792	26.286
Random data	10,240	12,288	112,044	0.439
Centromere	11,001	12,800	35,200	1.454
Cat-big.jpg	105,359	107,008	1,082,584	0.395

Table 5.4: Data before and after compression

Parameter	Value	Explanation
Sequencer	HiSeq 2500	Sequencing equipment to be simulated; we chose a commonly used Illumina sequencer
Read length	150	Length of simulated reads
Read depth	1, 2, 5, 10, 50	How many reads to collect for a given position. Higher numbers represent deeper sequencing (with more reads covering the same sequence).

Table 5.5: ART simulator parameters

Study	Code	Partition	Data encoded	Shannon information (bits)	Total bases	Rate	Error tolerance
Bancroft 2001	Triplets of A, C, T to uppercase letters	Address based, with indexes also concatenated on a dedicated key molecule	107 character English phrase	509	479	0.531	—
Church 2012	2 bases mapped to each binary digit	Address based	Text, images and source code	5,270,000	8,728,782	0.302	12 of 22 corrected
Goldman 2013	Huffman encoded, bases mapped to ternary digits and rotated	Purely address based assembly, fourfold redundancy due to overlaps between pieces	Text, PDF, audio, images	5,165,800	17,940,195	0.144	Recovered from 0.4%
Grass 2015	Based on Reed-Solomon code	Address based	Text	664,000	788,578	0.421	Recovered from 1%
Yazdi 2015	Word 6-tuple mapped to binary numbers and DNA sequences	Address based	Text	23,196	32,000	0.362	—
This study	Block cipher based on Levenshtein distances	Purely overlap based	Text, images, video, random data	83,824	111,192	0.377	Recovered from 1%

Table 5.6: Comparison of key publications

Chapter 6

Conclusions and Future Directions

Yeast has long been among the most powerful tools in experimental biology. In this thesis, I have described my attempts to push the boundaries of investigatory power. I have done so in three main respects:

- By constructing and characterizing extensive humanizations, ecolizations and plantizations of yeast I demonstrated the rapid applications of this technique to investigate hypotheses about basic biology. These strains have revealed important clues about the history of life on our planet (such as the evolution of localization). They have also vindicated the practical promise of humanized yeast: The idea that despite being a very distant evolutionary cousin, yeast can still have genetic analogs for human molecular function and dysfunction. The pink cell phenotype observed in humanized and ecolized yeast iron chelatases is striking in its resemblance to human porphyria, and an excellent example of how humanized can be used to assay disease alleles and functional epistasis.
- By developing a practical, rapid method of ortholog swapping and combination, I provided the tools for extension and generalization of the humanized yeast idea to many other pathways and genes. The excit-

ing practical benefit of humanized yeast strains is that they are a very practical chassis for personalized medicine and on-demand profiling of new human allelic variants without laborious human cell experiments. Straightforward, streamlined experimental tools are vital if this is to become a reality.

- By describing a method for digital information storage in yeast, I laid the groundwork for expanding the use of yeast beyond even biology, into information technology. Biological computing has long been a goal of synthetic biology[152], but so far biological computing devices have lagged decades behind electronic ones in terms of scale and speed. Although biological circuits attract much attention, it may very well be that the very first union of biology and computing will be in the realm of data storage, not processing. At any rate, the idea of DNA as a digital archival medium has already managed to capture the imaginations of industry researchers as well as academics[129].

An insightful review by Laurent and colleagues[153], frames the concept of humanization in terms of degrees. The lowest, degree 0, is the study of yeast which has not been humanized at all. Degree 1 is simply expressing human genes in yeast with no regard for orthology or complementation, a classic strategy of molecular biology. Degree 2 is humanization of specific positions within genes. Degree 3 is humanization of individual genes[8, 37]. The highest, degree 4, is humanization of entire pathways. At the time, few studies directly addressed this last, most ambitious category. The work described in Chapter

4 represents an early foray into humanization of the highest degree.

Having now arrived at the terminus of this taxonomy, I would like to propose a slightly different view to complement that of Laurent et al.: The notion of "resolution" in humanization. All humanization involves replacing *units* of the yeast genome with corresponding (often orthologous) units of the human genome (Figure 6.2). In the classic case, the units are genes, or rather coding sequences. Therefore, there is an implicit assumption in this approach that the salient unit of evolutionary selection is the gene. By going further and considering the swapping of whole pathways and complexes, we can access an alternative context where the evolution is thought to act on the genetic module. Indeed, [8] shows that there are two different classes of modules with regards to how evolutionary forces act on them. In one case, the module is restricted from divergence, while in the other it is not. With pathway-level humanization, we can query high-order genetic mechanisms, such as epistasis and regulation within the pathway. However, another interesting avenue of research would be to swap the yeast pathway with a hybrid one, which is a mixture of genes from humans and a third species. This would enable yeast to serve as a proxy for querying ortholog swapping from human to any other species while retaining the ease of yeast experimentation.

In the opposite direction, it is possible to swap smaller units, such as individual protein domains. Generalizing this, it would not be difficult to generate chimeras of yeast and human genes[154]. By surveying several chimeras of the same gene, we can obtain residue-level information of swappability (Fig-

ure 6.1). This information could potentially reveal foci of divergence: Individual sites that account for much of the fitness cost or non-swappability of a given ortholog. With chimeras of human alleles, human variation could also be explored at a precise level.

The unique biology of yeast makes it a very tractable organism for the experimentalist. Indeed, it must surely be one of the most studied organisms on the planet. One might even wonder if yeast research is close to being "done". With the advent of CRISPR, other model organisms seem poised to close this gap and become tractable to the same degree. However, those same technologies also boost the utility of yeast even further. A wealth of high-throughput studies has developed staggering know-how and infrastructure for investigation of global hypothesis about the whole of yeast genome and proteome. With ortholog swapping, this power can be leveraged to gain knowledge on the biology of humans and many other organisms. As we complete our understanding of yeast as an organism of study, we enter the era of yeast as a tremendous platform for the study of other organisms.

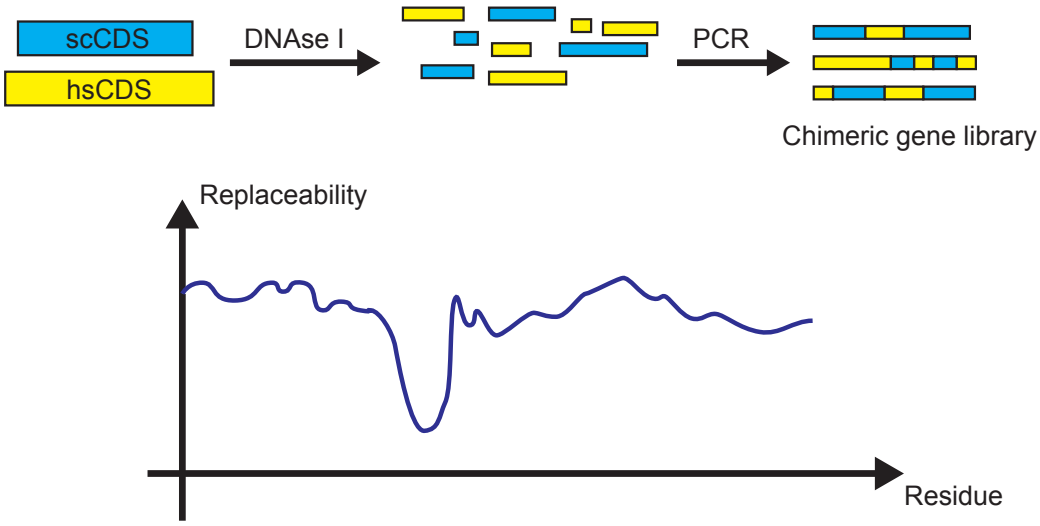

Figure 6.1: DNA shuffling can be used to gain residue-resolution of divergence. Many yeast/human chimeras are generated by mixing pieces of yeast and human sequence with PCR. Each chimera is then used to replace the yeast gene and the strain is characterized (such as by measuring growth rate). The results can be synthesized to find those residues which when humanized, cause the most significant impairment. For poor complementors, these residues could account for the majority of the barrier to swappability, and indicate divergent vs. conserved regions of the gene.

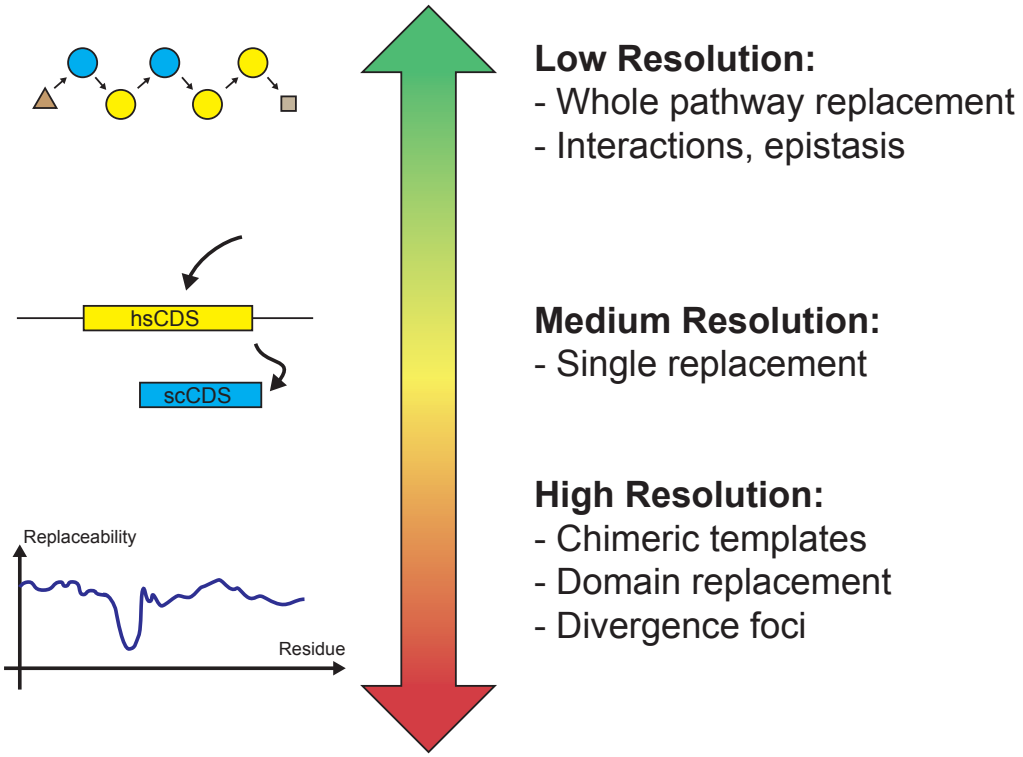

Figure 6.2: All humanization involves replacing *units* of the yeast genome with corresponding (often orthologous) units of the human genome. Traditionally, the unit is a gene. More precisely, in our work, it has been the coding sequence of each gene. Going to a "higher" degree affords a broader look and provides information about epistasis. The unit is then the "module": A pathway or protein complex. However, it is also possible to examine smaller units — these include the individual sites of degree 2, but also humanization of domains and introduction of chimeras of human and yeast sequence. A systematic survey of different chimeras for the same gene could provide a residue-level readout of replaceability, and potentially identify foci of divergence within the gene. Chimerizing human alleles could permit deep functional scans of human polymorphism.

Bibliography

[1] Samuel. "Investigation of Ancient Egyptian Baking and Brewing Methods by Correlative Microscopy". In: *Science (New York, N.Y.)* 273 (5274 July 1996), pp. 488–490. ISSN: 1095-9203.

[2] Diego Libkind et al. "Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast." In: *Proceedings of the National Academy of Sciences of the United States of America* 108 (35 Aug. 2011), pp. 14539–14544. ISSN: 1091-6490. DOI: 10.1073/pnas.1105430108.

[3] Delphine Sicard and Jean-Luc Legras. "Bread, beer and wine: yeast domestication in the Saccharomyces sensu stricto complex." In: *Comptes rendus biologies* 334 (3 Mar. 2011), pp. 229–236. ISSN: 1768-3238. DOI: 10.1016/j.crvi.2010.12.016.

[4] Bruno Müller and Ueli Grossniklaus. "Model organisms–A historical perspective." In: *Journal of proteomics* 73 (11 Oct. 2010), pp. 2054–2063. ISSN: 1876-7737. DOI: 10.1016/j.jprot.2010.08.002.

[5] Francesca Storici and Michael A Resnick. "The delitto perfetto approach to in vivo site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast." In: *Methods in enzymology* 409 (2006), pp. 329–345. ISSN: 0076-6879. DOI: 10.1016/S0076-6879(05)09019-1.

[6] Guri Giaever et al. "Functional profiling of the Saccharomyces cerevisiae genome." In: *Nature* 418 (6896 July 2002), pp. 387–391. ISSN: 0028-0836. DOI: 10.1038/nature00935.

[7] Maureen E Hillenmeyer et al. "The chemical genomic portrait of yeast: uncovering a phenotype for all genes." In: *Science (New York, N.Y.)* 320 (5874 Apr. 2008), pp. 362–365. ISSN: 1095-9203. DOI: 10.1126/science.1150021.

[8] Aashiq H Kachroo et al. "Systematic humanization of yeast genes reveals conserved functions and genetic modularity." In: *Science (New York, N.Y.)* 348 (6237 May 2015), pp. 921–925. ISSN: 1095-9203. DOI: 10.1126/science.aaa0769.

[9] S Fields and O Song. "A novel genetic system to detect protein-protein interactions." In: *Nature* 340 (6230 July 1989), pp. 245–246. ISSN: 0028-0836. DOI: 10.1038/340245a0.

[10] P Uetz et al. "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." In: *Nature* 403 (6770 Feb. 2000), pp. 623–627. ISSN: 0028-0836. DOI: 10.1038/35001009.

[11] A H Tong et al. "Systematic genetic analysis with ordered arrays of yeast deletion mutants." In: *Science (New York, N.Y.)* 294 (5550 Dec. 2001), pp. 2364–2368. ISSN: 0036-8075. DOI: 10.1126/science.1065810.

[12] Michael Costanzo et al. "The Genetic Landscape of a Cell". In: *Science* 327.5964 (2010), pp. 425–431. ISSN: 0036-8075. DOI: 10.1126/science.1180823. eprint: https://science.sciencemag.org/content/327/5964/425.full.pdf. URL: https://science.sciencemag.org/content/327/5964/425.

[13] Luciano A Marraffini and Erik J Sontheimer. "CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea." In: *Nature reviews. Genetics* 11 (3 Mar. 2010), pp. 181–190. ISSN: 1471-0064. DOI: 10.1038/nrg2749.

[14] Devaki Bhaya, Michelle Davison, and Rodolphe Barrangou. "CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation." In: *Annual review of genetics* 45 (2011), pp. 273–297. ISSN: 1545-2948. DOI: 10.1146/annurev-genet-110410-132430.

[15] R J Roberts. "Restriction endonucleases." In: *CRC critical reviews in biochemistry* 4 (2 Nov. 1976), pp. 123–164. ISSN: 0045-6411.

[16] David A Wright et al. "TALEN-mediated genome editing: prospects and perspectives." In: *The Biochemical journal* 462 (1 Aug. 2014), pp. 15–24. ISSN: 1470-8728. DOI: 10.1042/BJ20140295.

[17] Fyodor D Urnov et al. "Genome editing with engineered zinc finger nucleases." In: *Nature reviews. Genetics* 11 (9 Sept. 2010), pp. 636–646. ISSN: 1471-0064. DOI: 10.1038/nrg2842.

[18] Martin Jinek et al. "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity." In: *Science (New York, N.Y.)* 337 (6096 Aug. 2012), pp. 816–821. ISSN: 1095-9203. DOI: 10.1126/science.1225829.

[19] Shengdar Q Tsai et al. "Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing." In: *Nature biotechnology* 32 (6 June 2014), pp. 569–576. ISSN: 1546-1696. DOI: 10.1038/nbt.2908.

[20] James E DiCarlo et al. "Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems." In: *Nucleic acids research* 41 (7 Apr. 2013), pp. 4336–4343. ISSN: 1362-4962. DOI: 10.1093/nar/gkt135.

[21] Azat Akhmetov et al. "Single-step Precision Genome Editing in Yeast Using CRISPR-Cas9." In: *Bio-protocol* 8 (6 Mar. 2018). ISSN: 2331-8325. DOI: 10.21769/BioProtoc.2765.

[22] Owen W Ryan et al. "Selection of chromosomal DNA libraries using a multiplex CRISPR system." In: *eLife* 3 (Aug. 2014). ISSN: 2050-084X. DOI: 10.7554/eLife.03703.

[23] Michael E Lee et al. "A highly characterized yeast toolkit for modular, multipart assembly". In: *ACS synthetic biology* 4.9 (2015), pp. 975–986.

[24] John G Doench et al. "Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9." In: *Nature biotechnology* 34 (2 Feb. 2016), pp. 184–191. ISSN: 1546-1696. DOI: 10.1038/nbt.3437.

[25] K A Nasmyth and S I Reed. "Isolation of genes by complementation in yeast: molecular cloning of a cell-cycle gene." In: *Proceedings of the National Academy of Sciences of the United States of America* 77 (4 Apr. 1980), pp. 2119–2123. ISSN: 0027-8424.

[26] T Toda et al. "In yeast, RAS proteins are controlling elements of adenylate cyclase." In: *Cell* 40 (1 Jan. 1985), pp. 27–36. ISSN: 0092-8674.

[27] M G Lee and P Nurse. "Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc2." In: *Nature* 327 (6117 1987), pp. 31–35. ISSN: 0028-0836. DOI: 10.1038/327031a0.

[28] Katrin Paeschke et al. "Pif1 family helicases suppress genome instability at G-quadruplex motifs." In: *Nature* 497 (7450 May 2013), pp. 458–462. ISSN: 1476-4687. DOI: 10.1038/nature12149.

[29] John O Woods et al. "Prediction of gene-phenotype associations in humans, mice, and plants using phenologs." In: *BMC bioinformatics* 14 (June 2013), p. 203. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-203.

[30] Kriston L McGary et al. "Systematic discovery of nonobvious human disease models through orthologous phenotypes." In: *Proceedings of the National Academy of Sciences of the United States of America* 107

(14 Apr. 2010), pp. 6544–6549. ISSN: 1091-6490. DOI: 10.1073/pnas.0910200107.

[31] Hye Ji Cha et al. "Evolutionarily repurposed networks reveal the well-known antifungal drug thiabendazole to be a novel vascular disrupting agent." In: *PLoS biology* 10 (8 2012), e1001379. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001379.

[32] Michael J Osborn and J Ross Miller. "Rescuing yeast mutants with human genes." In: *Briefings in functional genomics & proteomics* 6 (2 June 2007), pp. 104–111. ISSN: 1473-9550. DOI: 10.1093/bfgp/elm017.

[33] Akil Hamza et al. "Complementation of Yeast Genes with Human Genes as an Experimental Platform for Functional Testing of Human Genetic Variants." eng. In: *Genetics* 201 (3 Nov. 2015), pp. 1263–74.

[34] Song Sun et al. "An extended set of yeast-based functional assays accurately identifies human disease mutations." eng. In: *Genome research* 26 (5 May 2016), pp. 670–80.

[35] Kevin P. O'Brien, Maido Remm, and Erik L. L. Sonnhammer. "Inparanoid: a comprehensive database of eukaryotic orthologs." eng. In: *Nucleic acids research* 33 (Database issue Jan. 2005), pp. D476–80.

[36] Fan Yang et al. "Identifying pathogenicity of human variants via paralog-based yeast complementation." In: *PLoS genetics* 13 (5 May 2017), e1006779. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1006779.

[37] Aashiq H Kachroo et al. "Systematic bacterialization of yeast genes identifies a near-universally swappable pathway". In: *eLife* 6 (2017).

[38] Michael A. Cusumano, Yiorgos Mylonadis, and Richard S. Rosenbloom. "Strategic Maneuvering and Mass-Market Dynamics: The Triumph of VHS over Beta". In: *Business History Review* 66.1 (1992), pp. 51–94. DOI: 10.2307/3117053.

[39] J. R. Brown and W. F. Doolittle. "Archaea and the prokaryote-to-eukaryote transition." eng. In: *Microbiology and molecular biology reviews : MMBR* 61 (4 Dec. 1997), pp. 456–502.

[40] W. Martin and M. Muller. "The hydrogen hypothesis for the first eukaryote." eng. In: *Nature* 392 (6671 Mar. 1998), pp. 37–41.

[41] Douglas L. Theobald. "A formal test of the theory of universal common ancestry." eng. In: *Nature* 465 (7295 May 2010), pp. 219–22.

[42] Toni Gabaldon and Eugene V. Koonin. *Functional and evolutionary implications of gene orthology.* eng. England, May 2013.

[43] David Lee, Oliver Redfern, and Christine Orengo. "Predicting protein function from sequence and structure." eng. In: *Nature reviews. Molecular cell biology* 8 (12 Dec. 2007), pp. 995–1005.

[44] J. Michael Cherry et al. "Saccharomyces Genome Database: the genomics resource of budding yeast." eng. In: *Nucleic acids research* 40 (Database issue Jan. 2012), pp. D700–5.

[45] Sven Heinicke et al. "The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists." eng. In: *PloS one* 2 (8 Aug. 2007). NLM: Original DateCompleted: 20070828, e766.

[46] M. Bulmer. "The selection-mutation-drift theory of synonymous codon usage." eng. In: *Genetics* 129 (3 Nov. 1991), pp. 897–907.

[47] P. M. Sharp et al. "Codon usage: mutational bias, translational selection, or both?" eng. In: *Biochemical Society transactions* 21 (4 Nov. 1993), pp. 835–41.

[48] Oliver Jardine et al. "Comparison of the small molecule metabolic enzymes of Escherichia coli and Saccharomyces cerevisiae." eng. In: *Genome research* 12 (6 June 2002), pp. 916–29.

[49] Jose Manuel Peregrin-Alvarez, Sophia Tsoka, and Christos A. Ouzounis. "The phylogenetic extent of metabolic enzymes and pathways." eng. In: *Genome research* 13 (3 Mar. 2003), pp. 422–7.

[50] F. R. Blattner et al. "The complete genome sequence of Escherichia coli K-12." eng. In: *Science (New York, N.Y.)* 277 (5331 Sept. 1997), pp. 1453–62.

[51] M. Ashburner et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." eng. In: *Nature genetics* 25 (1 May 2000), pp. 25–9.

[52] M. Kanehisa and S. Goto. "KEGG: kyoto encyclopedia of genes and genomes." eng. In: *Nucleic acids research* 28 (1 Jan. 2000), pp. 27–30.

[53] Mridusmita Saikia et al. "Codon optimality controls differential mRNA translation during amino acid starvation." eng. In: *RNA (New York, N.Y.)* 22 (11 Nov. 2016), pp. 1719–1727.

[54] Ilka U. Heinemann, Martina Jahn, and Dieter Jahn. "The biochemistry of heme biosynthesis." eng. In: *Archives of biochemistry and biophysics* 474 (2 June 2008), pp. 238–51.

[55] Liang Yin and Carl E. Bauer. "Controlling the delicate balance of tetrapyrrole biosynthesis." eng. In: *Philosophical transactions of the*

Royal Society of London. Series B, Biological sciences 368 (1622 July 2013), p. 20120262.

[56] Cheuk Hei Ho et al. "A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds." eng. In: *Nature biotechnology* 27 (4 Apr. 2009), pp. 369–77.

[57] J. J. Mukherjee and E. E. Dekker. "2-Amino-3-ketobutyrate CoA ligase of Escherichia coli: stoichiometry of pyridoxal phosphate binding and location of the pyridoxyllysine peptide in the primary structure of the enzyme." eng. In: *Biochimica et biophysica acta* 1037 (1 Jan. 1990), pp. 24–9.

[58] UniProt Consortium. "UniProt: a hub for protein information." In: *Nucleic acids research* 43 (Database issue Jan. 2015), pp. D204–D212. ISSN: 1362-4962. DOI: 10.1093/nar/gku989.

[59] L. L. Ilag and D. Jahn. "Activity and spectroscopic properties of the Escherichia coli glutamate 1-semialdehyde aminotransferase and the putative active site mutant K265R." eng. In: *Biochemistry* 31 (31 Aug. 1992), pp. 7143–51.

[60] Stefan Schauer et al. "Escherichia coli glutamyl-tRNA reductase. Trapping the thioester intermediate." eng. In: *The Journal of biological chemistry* 277 (50 Dec. 2002), pp. 48657–63.

[61] Judice L. Y. Koh et al. "CYCLoPs: A Comprehensive Database Constructed from Automated Analysis of Protein Abundance and Subcellular Localization Patterns in Saccharomyces cerevisiae." eng. In: *G3 (Bethesda, Md.)* 5 (6 Apr. 2015), pp. 1223–32.

[62] Malvina Papanastasiou et al. "The Escherichia coli peripheral inner membrane proteome." eng. In: *Molecular & cellular proteomics : MCP* 12 (3 Mar. 2013), pp. 599–610.

[63] J. Bloomer et al. "Molecular defects in ferrochelatase in patients with protoporphyria requiring liver transplantation." eng. In: *The Journal of clinical investigation* 102 (1 July 1998), pp. 107–14.

[64] W. E. Schauer and J. R. Mattoon. "Heterologous expression of human 5-aminolevulinate dehydratase in Saccharomyces cerevisiae." eng. In: *Current genetics* 17 (1 Jan. 1990), pp. 1–6.

[65] Nobuyoshi Mochizuki et al. "The cell biology of tetrapyrroles: a life and death struggle." eng. In: *Trends in plant science* 15 (9 Sept. 2010), pp. 488–98.

[66] A. G. Smith et al. “Isolation of a cDNA encoding chloroplast ferrochelatase from Arabidopsis thaliana by functional complementation of a yeast mutant.” eng. In: *The Journal of biological chemistry* 269 (18 May 1994), pp. 13405–13.

[67] L. L. Ilag, A. M. Kumar, and D. Soll. “Light regulation of chlorophyll biosynthesis at the level of 5-aminolevulinate formation in Arabidopsis.” eng. In: *The Plant cell* 6 (2 Feb. 1994), pp. 265–75.

[68] Ryouichi Tanaka, Koichi Kobayashi, and Tatsuru Masuda. “Tetrapyrrole Metabolism in Arabidopsis thaliana.” eng. In: *The arabidopsis book* 9 (2011), e0145.

[69] Jutta Papenbrock et al. “Expression studies in tetrapyrrole biosynthesis: inverse maxima of magnesium chelatase and ferrochelatase activity during cyclic photoperiods”. In: *Planta* 208.2 (1999), pp. 264–273.

[70] G. C. Ferreira et al. “Organization of the terminal two enzymes of the heme biosynthetic pathway. Orientation of protoporphyrinogen oxidase and evidence for a membrane complex.” eng. In: *The Journal of biological chemistry* 263 (8 Mar. 1988), pp. 3835–9.

[71] B. Grandchamp, N. Phung, and Y. Nordmann. “The mitochondrial localization of coproporphyrinogen III oxidase.” eng. In: *The Biochemical journal* 176 (1 Oct. 1978), pp. 97–102.

[72] Ludek Koreny and Miroslav Obornik. *Sequence evidence for the presence of two tetrapyrrole pathways in Euglena gracilis.* eng. England, 2011.

[73] Miroslav Obornik and Beverley R. Green. “Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes.” eng. In: *Molecular biology and evolution* 22 (12 Dec. 2005), pp. 2343–53.

[74] T. A. Dailey and H. A. Dailey. “Human protoporphyrinogen oxidase: expression, purification, and characterization of the cloned enzyme.” eng. In: *Protein science : a publication of the Protein Society* 5 (1 Jan. 1996), pp. 98–105.

[75] I. Lermontova et al. “Cloning and characterization of a plastidal and a mitochondrial isoform of tobacco protoporphyrinogen IX oxidase.” eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 94 (16 Aug. 1997), pp. 8895–900.

[76] S. Narita et al. “Molecular cloning and characterization of a cDNA that encodes protoporphyrinogen oxidase of Arabidopsis thaliana.” eng. In: *Gene* 182 (1-2 Dec. 1996), pp. 169–75.

[77] H. J. Lee et al. "Transgenic rice plants expressing a Bacillus subtilis protoporphyrinogen oxidase gene are resistant to diphenyl ether herbicide oxyfluorfen." eng. In: *Plant & cell physiology* 41 (6 June 2000), pp. 743–9.

[78] I. King Jordan et al. "Essential genes are more evolutionarily conserved than are nonessential genes in bacteria." eng. In: *Genome research* 12 (6 June 2002), pp. 962–8.

[79] Zhi Wang and Jianzhi Zhang. "Why is the correlation between gene importance and gene evolutionary rate so weak?" eng. In: *PLoS genetics* 5 (1 Jan. 2009), e1000329.

[80] R. Jain, M. C. Rivera, and J. A. Lake. "Horizontal gene transfer among genomes: the complexity hypothesis." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 96 (7 Mar. 1999), pp. 3801–6.

[81] Olof Emanuelsson et al. "Locating proteins in the cell using TargetP, SignalP and related tools." eng. In: *Nature protocols* 2 (4 2007), pp. 953–71.

[82] Zhijian Li et al. "Systematic exploration of essential yeast gene function with temperature-sensitive mutants." eng. In: *Nature biotechnology* 29 (4 Apr. 2011), pp. 361–7.

[83] Xuewen Pan et al. "A robust toolkit for functional profiling of the yeast genome." eng. In: *Molecular cell* 16 (3 Nov. 2004), pp. 487–96.

[84] Erik L. L. Sonnhammer and Gabriel Ostlund. "InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic." eng. In: *Nucleic acids research* 43 (Database issue Jan. 2015), pp. D234–9.

[85] Jaime Huerta-Cepas et al. "eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences." eng. In: *Nucleic acids research* 44 (D1 Jan. 2016), pp. D286–93.

[86] Adrian M. Altenhoff et al. "The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements." eng. In: *Nucleic acids research* 43 (Database issue Jan. 2015), pp. D240–9.

[87] Jindan Zhou and Kenneth E. Rudd. "EcoGene 3.0." eng. In: *Nucleic acids research* 41 (Database issue Jan. 2013), pp. D613–24.

[88] Nils A. Kulak et al. “Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells.” eng. In: *Nature methods* 11 (3 Mar. 2014), pp. 319–24.

[89] Nicholas T. Ingolia et al. “Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.” eng. In: *Science (New York, N.Y.)* 324 (5924 Apr. 2009), pp. 218–23.

[90] L. Arike et al. “Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli.” eng. In: *Journal of proteomics* 75 (17 Sept. 2012), pp. 5437–48.

[91] Chris Stark et al. “BioGRID: a general repository for interaction datasets.” eng. In: *Nucleic acids research* 34 (Database issue Jan. 2006), pp. D535–9.

[92] Eibe Frank et al. “Data mining in bioinformatics using Weka.” eng. In: *Bioinformatics (Oxford, England)* 20 (15 Oct. 2004), pp. 2479–81.

[93] J. Bassel et al. “Mutant of the yeast Saccharomycopsis lipolytica that accumulates and excretes protorphyrin IX.” eng. In: *Journal of bacteriology* 123 (1 July 1975), pp. 118–22.

[94] Theresa P Pretlow and Fred Sherman. “Porphyrins and zinc porphyrins in normal and mutant strains of yeast”. In: *Biochimica Et Biophysica Acta (BBA)-General Subjects* 148.3 (1967), pp. 629–644.

[95] Ki-Min Bark et al. “Physicochemical Properties of Protoporphyrin IX by metal ions in Acetonitrile-Water Mixture solution”. In: *Bulletin of the Korean Chemical Society* 31.6 (2010), pp. 1633–1637.

[96] M. Kearse et al. “Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data”. In: *Bioinformatics* 28 (2012). DOI: 10.1093/bioinformatics/bts199.

[97] D Jahn, E Verkamp, and D Söll. “Glutamyl-transfer RNA: a precursor of heme and chlorophyll biosynthesis.” In: *Trends in biochemical sciences* 17 (6 June 1992), pp. 215–218. ISSN: 0968-0004.

[98] Premal Shah et al. “Rate-limiting steps in yeast protein translation.” In: *Cell* 153 (7 June 2013), pp. 1589–1601. ISSN: 1097-4172. DOI: 10.1016/j.cell.2013.05.049.

[99] J D Weinstein and S I Beale. “Separate physiological roles and subcellular compartments for two tetrapyrrole biosynthetic pathways in Euglena gracilis.” In: *The Journal of biological chemistry* 258 (11 June 1983), pp. 6799–6807. ISSN: 0021-9258.

[100] Sonja Storbeck et al. “A novel pathway for the biosynthesis of heme in Archaea: genome-based bioinformatic predictions and experimental evidence.” eng. In: *Archaea (Vancouver, B.C.)* 2010 (Dec. 2010), p. 175050.

[101] Florian Heigwer, Grainne Kerr, and Michael Boutros. “E-CRISP: fast CRISPR target site identification.” In: *Nature methods* 11 (2 Feb. 2014), pp. 122–123. ISSN: 1548-7105. DOI: 10.1038/nmeth.2812.

[102] W James Kent. “BLAT–the BLAST-like alignment tool.” In: *Genome research* 12 (4 Apr. 2002), pp. 656–664. ISSN: 1088-9051. DOI: 10.1101/gr.229202.

[103] Eleni P Mimitou and Lorraine S Symington. “DNA end resection: many nucleases make light work”. In: *DNA repair* 8.9 (2009), pp. 983–995.

[104] Xuefeng Chen et al. “Cell cycle regulation of DNA double-strand break end resection by Cdk1-dependent Dna2 phosphorylation”. In: *Nature structural & molecular biology* 18.9 (2011), pp. 1015–1019.

[105] J D Boeke et al. “5-Fluoroorotic acid as a selective agent in yeast molecular genetics.” In: *Methods in enzymology* 154 (1987), pp. 164–175. ISSN: 0076-6879.

[106] Cem Kuscu et al. “Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease.” In: *Nature Biotechnology* 32 (7 July 2014), pp. 677–683. ISSN: 1546-1696. DOI: 10.1038/nbt.2916.

[107] Xiaojun Xu, Dongsheng Duan, and Shi-Jie Chen. “CRISPR-Cas9 cleavage efficiency correlates strongly with target-sgRNA folding stability: from physical mechanism to off-target assessment.” In: *Scientific reports* 7 (1 Mar. 2017), p. 143. ISSN: 2045-2322. DOI: 10.1038/s41598-017-00180-1.

[108] James E DiCarlo et al. “Safeguarding CRISPR-Cas9 gene drives in yeast.” In: *Nature biotechnology* 33 (12 Dec. 2015), pp. 1250–1255. ISSN: 1546-1696. DOI: 10.1038/nbt.3412.

[109] Elizabeth A. Winzeler et al. “Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis”. In: *Science* 285.5429 (1999), pp. 901–906. ISSN: 0036-8075. DOI: 10.1126/science.285.5429.901. eprint: https://science.sciencemag.org/content/285/5429/901.full.pdf. URL: https://science.sciencemag.org/content/285/5429/901.

[110] Michael Costanzo et al. "A global genetic interaction network maps a wiring diagram of cellular function". In: *Science* 353.6306 (2016). ISSN: 0036-8075. DOI: 10.1126/science.aaf1420. eprint: https://science.sciencemag.org/content/353/6306/aaf1420.full.pdf. URL: https://science.sciencemag.org/content/353/6306/aaf1420.

[111] H A Orr. "The population genetics of speciation: the evolution of hybrid incompatibilities." In: *Genetics* 139 (4 Apr. 1995), pp. 1805–1813. ISSN: 0016-6731.

[112] John J Welch. "Accumulating Dobzhansky-Muller incompatibilities: reconciling theory and data." In: *Evolution; international journal of organic evolution* 58 (6 June 2004), pp. 1145–1156. ISSN: 0014-3820.

[113] Yo Suzuki et al. "Knocking out multigene redundancies via cycles of sexual assortment and fluorescence selection." In: *Nature methods* 8 (2 Feb. 2011), pp. 159–164. ISSN: 1548-7105. DOI: 10.1038/nmeth.1550.

[114] Simone Lubrano et al. "Development of a yeast-based system to identify new hBRAFV600E functional interactors." In: *Oncogene* 38 (8 Feb. 2019), pp. 1355–1366. ISSN: 1476-5594. DOI: 10.1038/s41388-018-0496-5.

[115] T D Petes. "Yeast ribosomal DNA genes are located on chromosome XII." In: *Proceedings of the National Academy of Sciences of the United States of America* 76 (1 Jan. 1979), pp. 410–414. ISSN: 0027-8424.

[116] Stephen A James et al. "Repetitive sequence variation and dynamics in the ribosomal DNA array of Saccharomyces cerevisiae as revealed by whole-genome resequencing." In: *Genome research* 19 (4 Apr. 2009), pp. 626–635. ISSN: 1088-9051. DOI: 10.1101/gr.084517.108.

[117] M Johnston et al. "The nucleotide sequence of Saccharomyces cerevisiae chromosome XII." In: *Nature* 387 (6632 Suppl May 1997), pp. 87–90. ISSN: 0028-0836.

[118] Meysam Moghbeli et al. "Mutational analysis of uroporphyrinogen III cosynthase gene in Iranian families with congenital erythropoietic porphyria." In: *Molecular biology reports* 39 (6 June 2012), pp. 6731–6735. ISSN: 1573-4978. DOI: 10.1007/s11033-012-1497-z.

[119] Jordi To-Figueras et al. "ALAS2 acts as a modifier gene in patients with congenital erythropoietic porphyria." In: *Blood* 118 (6 Aug. 2011), pp. 1443–1451. ISSN: 1528-0020. DOI: 10.1182/blood-2011-03-342873.

[120] L Ni and M Snyder. "A genomic study of the bipolar bud site selection pattern in Saccharomyces cerevisiae." In: *Molecular biology of the cell* 12 (7 July 2001), pp. 2147–2170. ISSN: 1059-1524. DOI: 10.1091/mbc.12.7.2147.

[121] C B Brachmann et al. "Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications." In: *Yeast (Chichester, England)* 14 (2 Jan. 1998), pp. 115–132. ISSN: 0749-503X. DOI: 10.1002/(SICI)1097-0061(19980130)14:2<115::AID-YEA204>3.0.CO;2-2.

[122] Robert C Edgar. "MUSCLE: multiple sequence alignment with high accuracy and high throughput." In: *Nucleic acids research* 32 (5 2004), pp. 1792–1797. ISSN: 1362-4962. DOI: 10.1093/nar/gkh340.

[123] Robert C Edgar. "MUSCLE: a multiple sequence alignment method with reduced time and space complexity." In: *BMC bioinformatics* 5 (Aug. 2004), p. 113. ISSN: 1471-2105. DOI: 10.1186/1471-2105-5-113.

[124] C. Bancroft et al. "Long-term storage of information in DNA". In: *Science* 293 (2001). DOI: 10.1126/science.293.5536.1763c.

[125] G. M. Church, Y. Gao, and S. Kosuri. "Next-generation digital information storage in DNA". In: *Science* 337 (2012). DOI: 10.1126/science.1226355.

[126] N. Goldman et al. "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA". In: *Nature* 494 (2013). DOI: 10.1038/nature11875.

[127] S. M. H. T. Yazdi et al. "A rewritable, random-access DNA-based storage system". In: *Sci Rep* 5 (2015). DOI: 10.1038/srep14138.

[128] G. C. Smith et al. "Some possible codes for encrypting data in DNA". In: *Biotechnol Lett* 25 (2003). DOI: 10.1023/A:1024539608706.

[129] James Bornholt et al. "A DNA-Based Archival Storage System". In: *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*. ASPLOS '16. Atlanta, Georgia, USA: ACM, 2016, pp. 637–649. ISBN: 978-1-4503-4091-5. DOI: 10.1145/2872362.2872397.

[130] Morten E Allentoft et al. "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils." In: *Proceedings. Biological sciences*

279 (1748 Dec. 2012), pp. 4724–4733. ISSN: 1471-2954. DOI: 10.1098/rspb.2012.1745.

[131] E. Palkopoulou et al. "Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth". In: *Curr Biol* 25 (2015). DOI: 10.1016/j.cub.2015.04.007.

[132] Vladimir I Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet physics doklady*. Vol. 10. 8. 1966, pp. 707–710.

[133] T. Lindahl and B. Nyberg. "Rate of depurination of native deoxyribonucleic acid". In: *Biochemistry* 11 (1972). DOI: 10.1021/bi00769a018.

[134] E. Willerslev et al. "Ancient biomolecules from deep ice cores reveal a forested southern Greenland". In: *Science* 317 (2007). DOI: 10.1126/science.1141758.

[135] M. Hedstrom. "Digital preservation: a time bomb for digital libraries". In: *Comput Hum* 31 (1997). DOI: 10.1023/A:1000676723815.

[136] M. Irie and Y. Okino. "Statistical analysis of lifetime distribution for optical recordable disks". In: *Jpn J Appl Phys* 45 (2006). DOI: 10.1143/JJAP.45.1460.

[137] O. Slattery et al. "Stability comparison of recordable optical discs-a study of error rates in harsh conditions". In: *J Res-Natl Inst Stand Technol* 109 (2004). DOI: 10.6028/jres.109.038.

[138] Z. Sun, J. Zhou, and R. Ahuja. "Unique melting behavior in phase-change materials for rewritable data storage". In: *Phys Rev Lett* 98 (2007). DOI: 10.1103/PhysRevLett.98.055505.

[139] T. Kuny. "The digital dark ages? Challenges in the preservation of electronic information". In: *Int Preserv News* 17 (1998).

[140] R. N. Grass et al. "Robust chemical preservation of digital information on DNA in silica with error-correcting codes". In: *Angew Chem (Int Ed Eng)* 54 (2015). DOI: 10.1002/anie.201411378.

[141] Florence Jessie MacWilliams and Neil James Alexander Sloane. *The theory of error-correcting codes*. Vol. 16. Elsevier, 1977.

[142] T. Buschmann and L. V. Bystrykh. "Levenshtein error-correcting barcodes for multiplexed DNA sequencing". In: *BMC bioinformatics* 14 (2013). DOI: 10.1186/1471-2105-14-272.

[143] L. V. Bystrykh. "Generalized DNA barcode design based on hamming codes". In: *PLoS One* 7 (2012). DOI: 10.1371/journal.pone.0036852.

[144] Q. Xu et al. "Design of 240,000 orthogonal 25mer DNA barcode probes". In: *Proc Natl Acad Sci* 106 (2009). DOI: 10.1073/pnas.0812506106.

[145] M. Hamady et al. "Error-correcting barcoded primers allow hundreds of samples to be pyrosequenced in multiplex". In: *Nat Methods* 5 (2008). DOI: 10.1038/nmeth.1184.

[146] D. W. Craig et al. "Identification of genetic variants using bar-coded multiplexed sequencing". In: *Nat Methods* 5 (2008). DOI: 10.1038/nmeth.1251.

[147] James A Brown et al. "Global analysis of gene function in yeast by quantitative phenotypic profiling." In: *Molecular systems biology* 2 (2006), p. 20060001. ISSN: 1744-4292. DOI: 10.1038/msb4100043.

[148] M. Hafner et al. "Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing". In: *Methods* 58 (2012). DOI: 10.1016/j.ymeth.2012.07.030.

[149] W. Huang et al. "ART: a next-generation sequencing read simulator". In: *Bioinformatics* 28 (2012). DOI: 10.1093/bioinformatics/btr708.

[150] Sriram Kosuri and George M Church. "Large-scale de novo DNA synthesis: technologies and applications". In: *Nature methods* 11.5 (2014), p. 499.

[151] Travis C Glenn. "Field guide to next-generation DNA sequencers." In: *Molecular ecology resources* 11 (5 Sept. 2011), pp. 759–769. ISSN: 1755-0998. DOI: 10.1111/j.1755-0998.2011.03024.x.

[152] Ahmad S Khalil and James J Collins. "Synthetic biology: applications come of age." In: *Nature reviews. Genetics* 11 (5 May 2010), pp. 367–379. ISSN: 1471-0064. DOI: 10.1038/nrg2775.

[153] Jon M Laurent et al. "Efforts to make and apply humanized yeast." In: *Briefings in functional genomics* 15 (2 Mar. 2016), pp. 155–163. ISSN: 2041-2657. DOI: 10.1093/bfgp/elv041.

[154] W P Stemmer. "Rapid evolution of a protein in vitro by DNA shuffling." In: *Nature* 370 (6488 Aug. 1994), pp. 389–391. ISSN: 0028-0836. DOI: 10.1038/370389a0.

[155] G. Sullivan and F. Weierud. "Breaking german army ciphers". In: *Cryptologia* 29 (2005). DOI: 10.1080/01611190508951299.

[156] Azat Akhmetov, Andrew D Ellington, and Edward M Marcotte. "A highly parallel strategy for storage of digital information in living cells."

In: *BMC biotechnology* 18 (1 Oct. 2018), p. 64. ISSN: 1472-6750. DOI: 10.1186/s12896-018-0476-4.

[157] Rodolphe Barrangou et al. "CRISPR provides acquired resistance against viruses in prokaryotes." In: *Science (New York, N.Y.)* 315 (5819 Mar. 2007), pp. 1709–1712. ISSN: 1095-9203. DOI: 10.1126/science.1138140.

[158] Jörg O Schulze et al. "Evolutionary relationship between initial enzymes of tetrapyrrole biosynthesis." In: *Journal of molecular biology* 358 (5 May 2006), pp. 1212–1220. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2006.02.064.

[159] E Verkamp et al. "Glutamyl-tRNA reductase from Escherichia coli and Synechocystis 6803. Gene structure and expression." In: *The Journal of biological chemistry* 267 (12 Apr. 1992), pp. 8275–8280. ISSN: 0021-9258.